Next Generation Sequencing: Introduction, Algorithms and Tools

Aik Choon Tan, Ph.D.
Associate Professor of Bioinformatics
Division of Medical Oncology
Department of Medicine
aikchoon.tan@ucdenver.edu

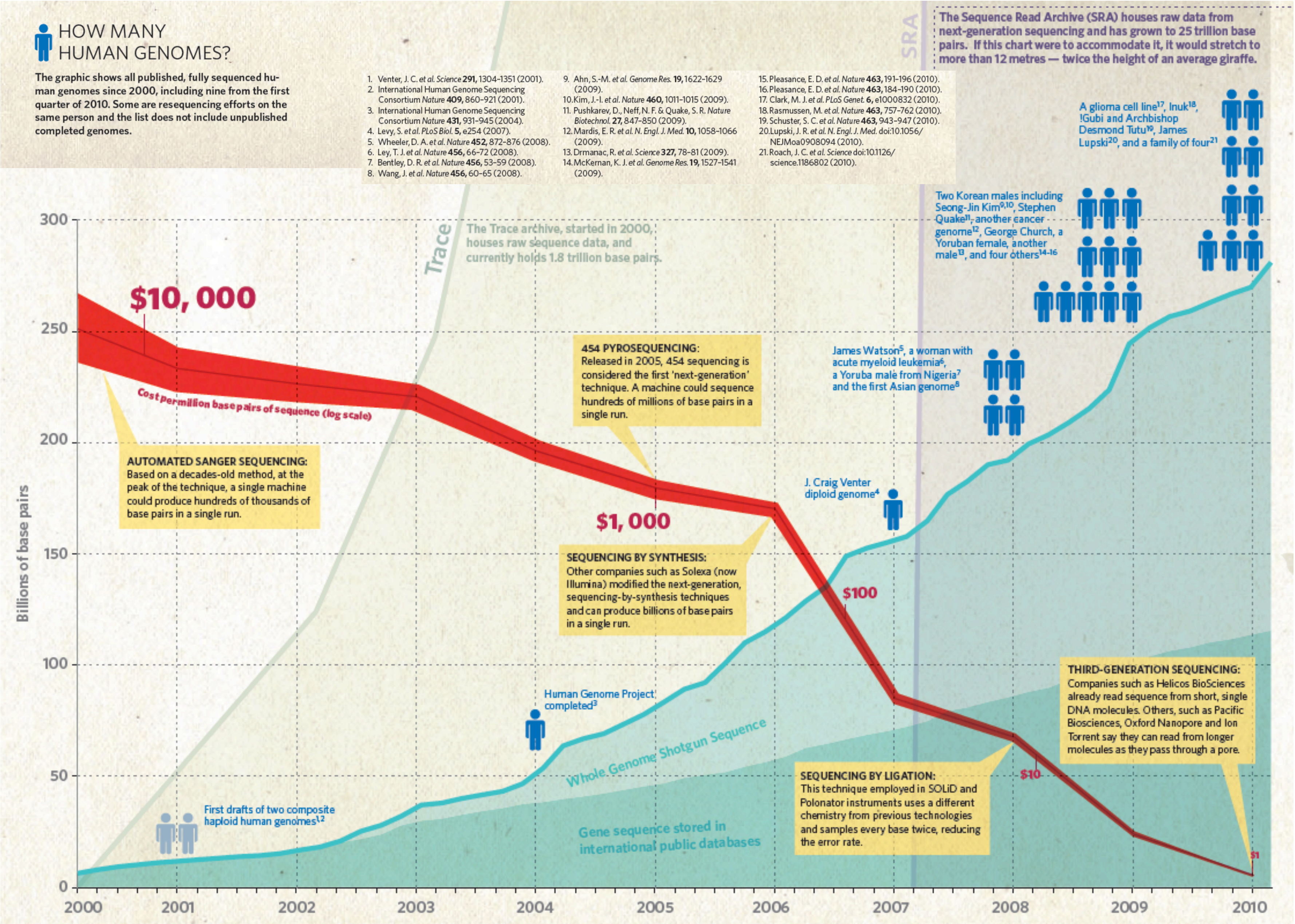
10/23/2018

http://tanlab.ucdenver.edu/labHomePage/teaching/CANB7640

Outline

- Introduction to Next Generation Sequencing Technologies
- Mapping Algorithm – Burrows-Wheeler Algorithm
- Tools to Analyze and Visualize NGS data

The Sequence Explosion



HOW MANY HUMAN GENOMES?

The graphic shows all published, fully sequenced human genomes since 2000, including nine from the first quarter of 2010. Some are resequencing efforts on the same person and the list does not include unpublished completed genomes.

1. Venter, J. C. *et al. Science* **291,** 1304–1351 (2001).
2. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
3. International Human Genome Sequencing Consortium *Nature* **431,** 931–945 (2004).
4. Levy, S. *et al. PLoS Biol.* **5,** e254 (2007).
5. Wheeler, D. A. *et al. Nature* **452,** 872–876 (2008).
6. Ley, T. J. *et al. Nature* **456,** 66–72 (2008).
7. Bentley, D. R. *et al. Nature* **456,** 53–59 (2008).
8. Wang, J. *et al. Nature* **456,** 60–65 (2008).
9. Ahn, S.-M. *et al. Genome Res.* **19,** 1622–1629 (2009).
10. Kim, J.-I. *et al. Nature* **460,** 1011–1015 (2009).
11. Pushkarev, D., Neff, N. F. & Quake, S. R. *Nature Biotechnol.* **27,** 847–850 (2009).
12. Mardis, E. R. *et al. N. Engl. J. Med.* **10,** 1058–1066 (2009).
13. Drmanac, R. *et al. Science* **327,** 78–81 (2009).
14. McKernan, K. J. *et al. Genome Res.* **19,** 1527–1541 (2009).
15. Pleasance, E. D. *et al. Nature* **463,** 191–196 (2010).
16. Pleasance, E. D. *et al. Nature* **463,** 184–190 (2010).
17. Clark, M. J. *et al. PLoS Genet.* **6,** e1000832 (2010).
18. Rasmussen, M. *et al. Nature* **463,** 757–762 (2010).
19. Schuster, S. C. *et al. Nature* **463,** 943–947 (2010).
20. Lupski, J. R. *et al. N. Engl. J. Med.* doi:10.1056/NEJMoa0908094 (2010).
21. Roach, J. C. *et al. Science* doi:10.1126/science.1186802 (2010).

The Sequence Read Archive (SRA) houses raw data from next-generation sequencing and has grown to 25 trillion base pairs. If this chart were to accommodate it, it would stretch to more than 12 metres — twice the height of an average giraffe.

The Trace archive, started in 2000, houses raw sequence data, and currently holds 1.8 trillion base pairs.

AUTOMATED SANGER SEQUENCING: Based on a decades-old method, at the peak of the technique, a single machine could produce hundreds of thousands of base pairs in a single run.

454 PYROSEQUENCING: Released in 2005, 454 sequencing is considered the first 'next-generation' technique. A machine could sequence hundreds of millions of base pairs in a single run.

SEQUENCING BY SYNTHESIS: Other companies such as Solexa (now Illumina) modified the next-generation, sequencing-by-synthesis techniques and can produce billions of base pairs in a single run.

SEQUENCING BY LIGATION: This technique employed in SOLiD and Polonator instruments uses a different chemistry from previous technologies and samples every base twice, reducing the error rate.

THIRD-GENERATION SEQUENCING: Companies such as Helicos BioSciences already read sequence from short, single DNA molecules. Others, such as Pacific Biosciences, Oxford Nanopore and Ion Torrent say they can read from longer molecules as they pass through a pore.

First drafts of two composite haploid human genomes[1,2]

Human Genome Project completed[3]

J. Craig Venter diploid genome[4]

James Watson[5], a woman with acute myeloid leukemia[6], a Yoruba male from Nigeria[7] and the first Asian genome[8]

Two Korean males including Seong-Jin Kim[9,10], Stephen Quake[11], another cancer genome[12], George Church, a Yoruban female, another male[13], and four others[14-16]

A glioma cell line[17], Inuk[18], !Gubi and Archbishop Desmond Tutu[19], James Lupski[20], and a family of four[21]

SRA
Trace
$10, 000
Cost per million base pairs of sequence (log scale)
$1, 000
$100
$10
$1
Whole Genome Shotgun Sequence
Gene sequence stored in international public databases
Billions of base pairs
300
250
200
150
100
50
0
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010

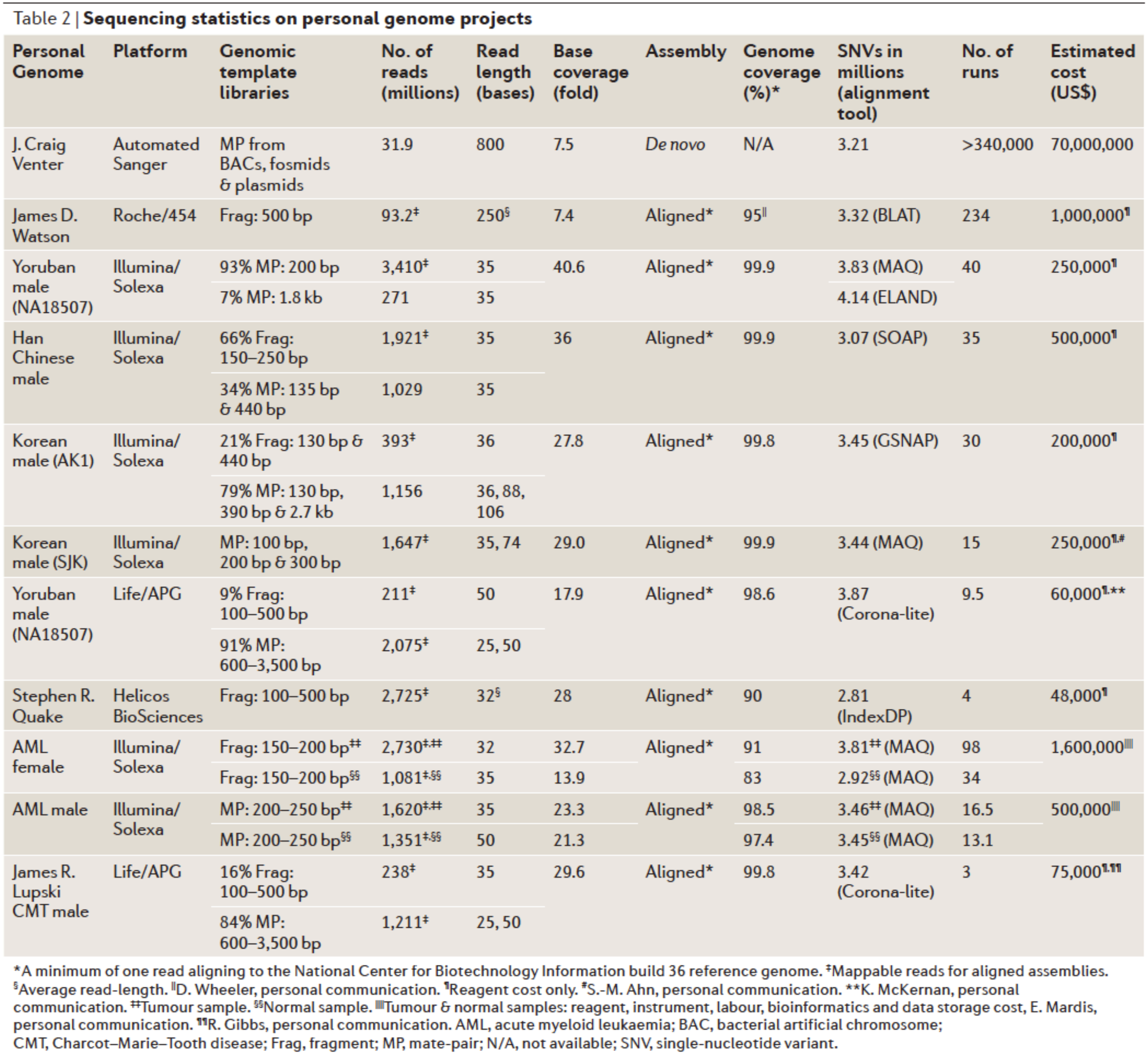


Table 2 | **Sequencing statistics on personal genome projects**

Personal Genome	Platform	Genomic template libraries	No. of reads (millions)	Read length (bases)	Base coverage (fold)	Assembly	Genome coverage (%)*	SNVs in millions (alignment tool)	No. of runs	Estimated cost (US$)
J. Craig Venter	Automated Sanger	MP from BACs, fosmids & plasmids	31.9	800	7.5	*De novo*	N/A	3.21	>340,000	70,000,000
James D. Watson	Roche/454	Frag: 500 bp	93.2‡	250§	7.4	Aligned*	95‖	3.32 (BLAT)	234	1,000,000¶
Yoruban male (NA18507)	Illumina/ Solexa	93% MP: 200 bp	3,410‡	35	40.6	Aligned*	99.9	3.83 (MAQ)	40	250,000¶
		7% MP: 1.8 kb	271	35				4.14 (ELAND)		
Han Chinese male	Illumina/ Solexa	66% Frag: 150–250 bp	1,921‡	35	36	Aligned*	99.9	3.07 (SOAP)	35	500,000¶
		34% MP: 135 bp & 440 bp	1,029	35						
Korean male (AK1)	Illumina/ Solexa	21% Frag: 130 bp & 440 bp	393‡	36	27.8	Aligned*	99.8	3.45 (GSNAP)	30	200,000¶
		79% MP: 130 bp, 390 bp & 2.7 kb	1,156	36, 88, 106						
Korean male (SJK)	Illumina/ Solexa	MP: 100 bp, 200 bp & 300 bp	1,647‡	35, 74	29.0	Aligned*	99.9	3.44 (MAQ)	15	250,000¶,#
Yoruban male (NA18507)	Life/APG	9% Frag: 100–500 bp	211‡	50	17.9	Aligned*	98.6	3.87 (Corona-lite)	9.5	60,000¶,**
		91% MP: 600–3,500 bp	2,075‡	25, 50						
Stephen R. Quake	Helicos BioSciences	Frag: 100–500 bp	2,725‡	32§	28	Aligned*	90	2.81 (IndexDP)	4	48,000¶
AML female	Illumina/ Solexa	Frag: 150–200 bp‡‡	2,730‡,‡‡	32	32.7	Aligned*	91	3.81‡‡ (MAQ)	98	1,600,000‖‖
		Frag: 150–200 bp§§	1,081‡,§§	35	13.9		83	2.92§§ (MAQ)	34	
AML male	Illumina/ Solexa	MP: 200–250 bp‡‡	1,620‡,‡‡	35	23.3	Aligned*	98.5	3.46‡‡ (MAQ)	16.5	500,000‖‖
		MP: 200–250 bp§§	1,351‡,§§	50	21.3		97.4	3.45§§ (MAQ)	13.1	
James R. Lupski CMT male	Life/APG	16% Frag: 100–500 bp	238‡	35	29.6	Aligned*	99.8	3.42 (Corona-lite)	3	75,000¶,¶¶
		84% MP: 600–3,500 bp	1,211‡	25, 50						

*A minimum of one read aligning to the National Center for Biotechnology Information build 36 reference genome. ‡Mappable reads for aligned assemblies. §Average read-length. ‖D. Wheeler, personal communication. ¶Reagent cost only. #S.-M. Ahn, personal communication. **K. McKernan, personal communication. ‡‡Tumour sample. §§Normal sample. ‖‖Tumour & normal samples: reagent, instrument, labour, bioinformatics and data storage cost, E. Mardis, personal communication. ¶¶R. Gibbs, personal communication. AML, acute myeloid leukaemia; BAC, bacterial artificial chromosome; CMT, Charcot–Marie–Tooth disease; Frag, fragment; MP, mate-pair; N/A, not available; SNV, single-nucleotide variant.

(Now, $1500 per genome)

APPLICATIONS OF NEXT-GENERATION SEQUENCING

Sequencing technologies — the next generation

Michael L. Metzker[*,‡]

Abstract | Demand has never been greater for revolutionary technologies that deliver fast, inexpensive and accurate genome information. This challenge has catalysed the development of next-generation sequencing (NGS) technologies. The inexpensive production of large volumes of sequence data is the primary advantage over conventional methods. Here, I present a technical review of template preparation, sequencing and imaging, genome alignment and assembly approaches, and recent advances in current and near-term commercially available NGS instruments. I also outline the broad range of applications for NGS technologies, in addition to providing guidelines for platform selection to address biological questions of interest.

Over the past four years, there has been a fundamental shift away from the application of automated Sanger sequencing for genome analysis. Prior to this departure, the automated Sanger method had dominated the industry for almost two decades and led to a number of monumental accomplishments, including the completion of the only finished-grade human genome sequence[1]. Despite many technical improvements during this era, the limitations of automated Sanger sequencing showed a need for new and improved technologies for sequencing large numbers of human genomes. Recent efforts have been directed towards the development of new methods, leaving Sanger sequencing with fewer reported advances. As such, automated Sanger sequencing is not covered here, and interested readers are directed to previous articles[2,3].

The automated Sanger method is considered as a 'first-generation' technology, and newer methods are referred to as next-generation sequencing (NGS). These newer technologies constitute various strategies that rely on a combination of template preparation, sequencing and imaging, and genome alignment and assembly methods. The arrival of NGS technologies in the marketplace has changed the way we think about scientific approaches in basic, applied and clinical research. In some respects, the potential of NGS is akin to the early days of PCR, with one's imagination being the primary limitation to its use. The major advance offered by NGS is the ability to produce an enormous volume of data cheaply — in some cases in excess of one billion short reads per instrument run. This feature expands the realm of experimentation beyond just determining the order of bases. For example, in gene-expression studies microarrays are now being replaced by seq-based methods, which can identify and quantify rare transcripts without prior knowledge of a particular gene and can provide information regarding alternative splicing and sequence variation in identified genes[4,5]. The ability to sequence the whole genome of many related organisms has allowed large-scale comparative and evolutionary studies to be performed that were unimaginable just a few years ago. The broadest application of NGS may be the resequencing of human genomes to enhance our understanding of how genetic differences affect health and disease. The variety of NGS features makes it likely that multiple platforms will coexist in the marketplace, with some having clear advantages for particular applications over others.

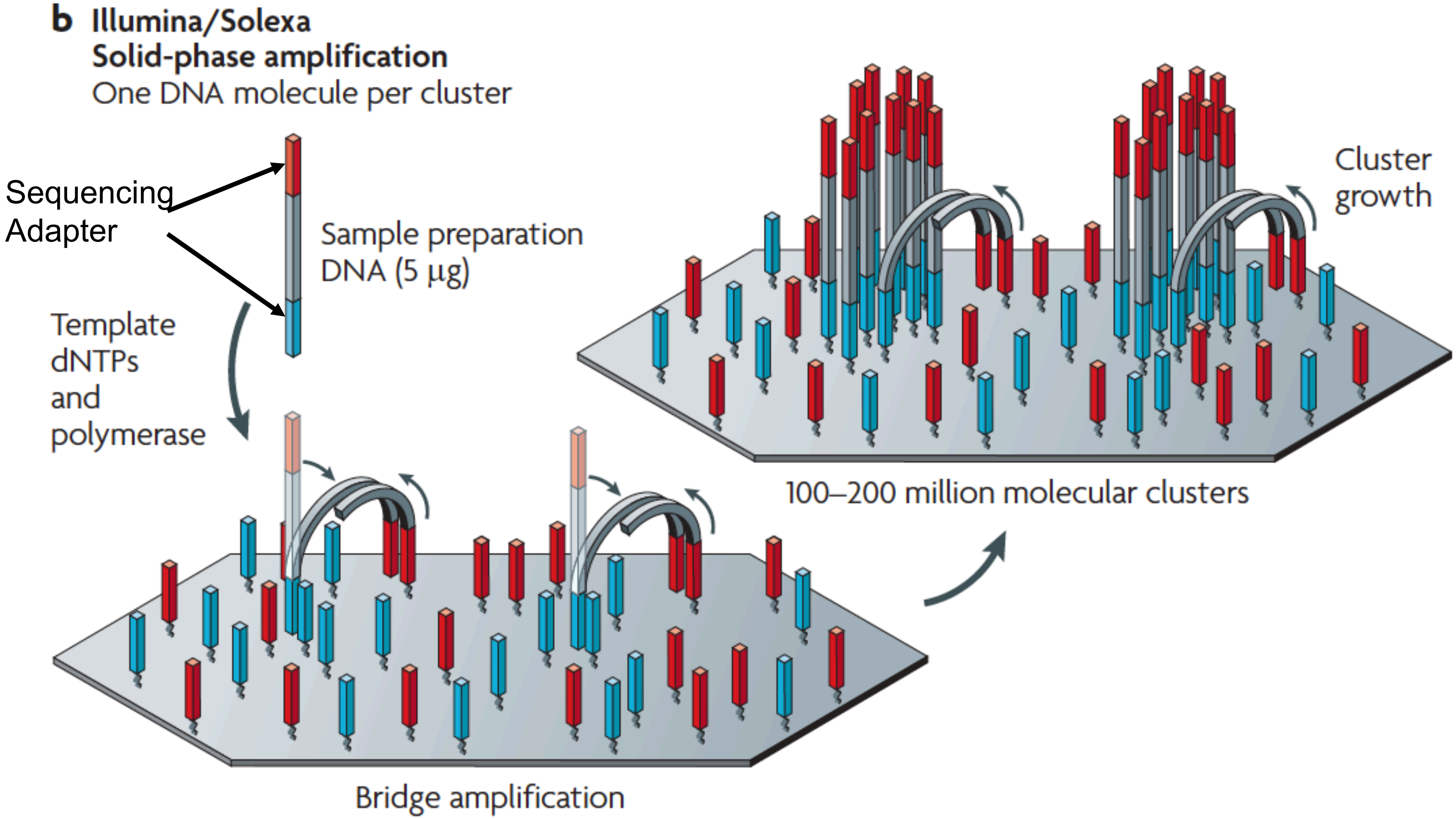
This Review focuses on commercially available technologies from Roche/454, Illumina/Solexa, Life/APG and Helicos BioSciences, the Polonator instrument and the near-term technology of Pacific Biosciences, who aim to bring their sequencing device to the market in 2010. Nanopore sequencing is not covered, although interested readers are directed to an article by Branton and colleagues[6], who describe the advances and remaining challenges for this technology. Here, I present a technical review of template preparation, sequencing and imaging, genome alignment and assembly, and current NGS platform performance to provide guidance on how these technologies work and how they may be applied to important biological questions. I highlight the applications of human genome resequencing using targeted and whole-genome approaches, and discuss the progress

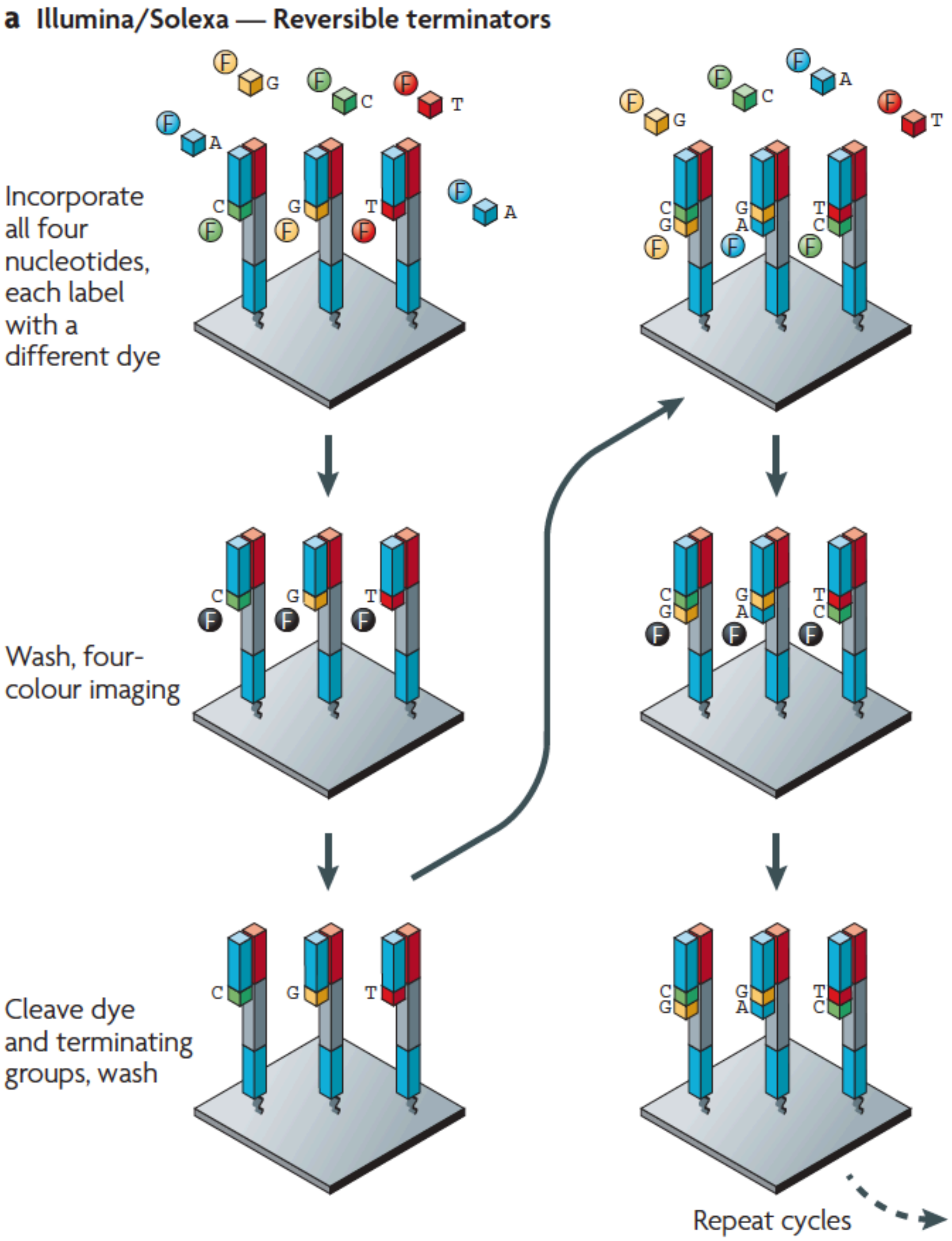
Automated Sanger sequencing
This process involves a mixture of techniques: bacterial cloning or PCR; template purification; labelling of DNA fragments using the chain termination method with energy transfer, dye-labelled dideoxynucleotides and a DNA polymerase; capillary electrophoresis; and fluorescence detection that provides four-colour plots to reveal the DNA sequence.

*Human Genome Sequencing Center and Department of Molecular & Human Genetics, Baylor College of Medicine, One Baylor Plaza, N1409, Houston, Texas 77030, USA.
‡LaserGen, Inc., 8052 El Rio Street, Houston, Texas 77054, USA.
e-mail: mmetzker@bcm.edu

doi:10.1038/nrg2626
Published online
8 December 2009

Illumina Next Generation Sequencing Technology (Sequencing-by-Synthesis)


b Illumina/Solexa
Solid-phase amplification
One DNA molecule per cluster
Sequencing Adapter
Sample preparation DNA (5 µg)
Template dNTPs and polymerase
Cluster growth
100–200 million molecular clusters
Bridge amplification



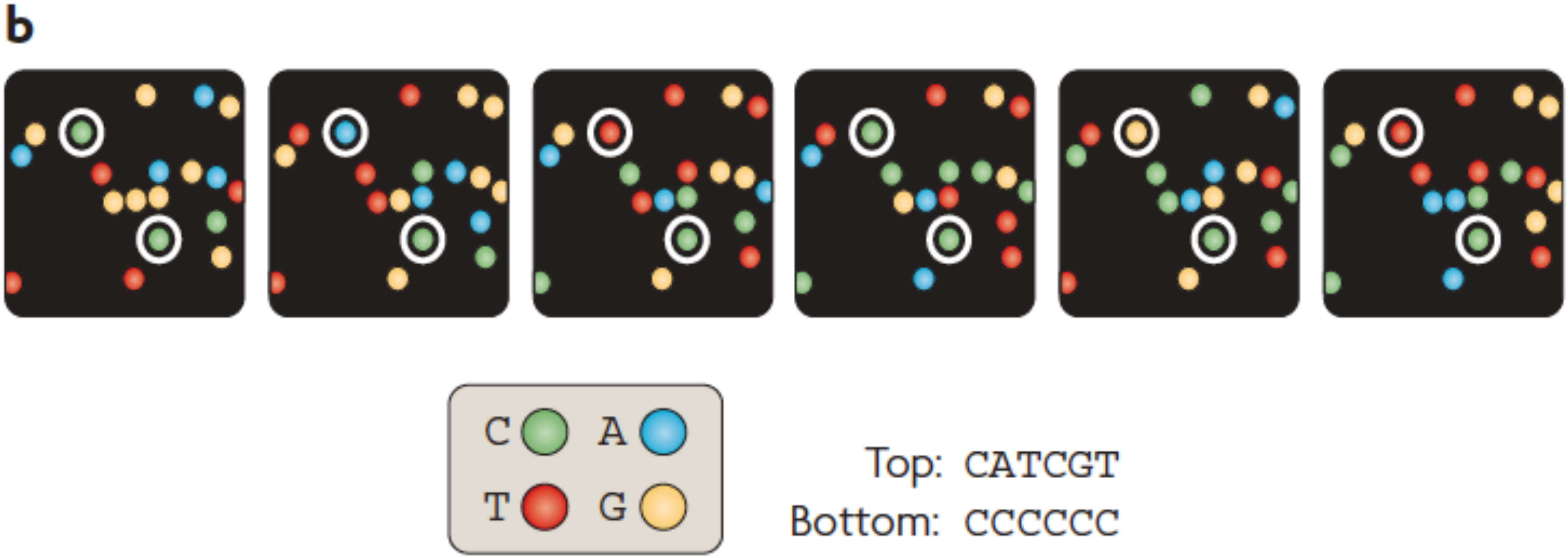
a Illumina/Solexa — Reversible terminators

Incorporate all four nucleotides, each label with a different dye

Wash, four-colour imaging

Cleave dye and terminating groups, wash

Repeat cycles


b
C
A
T
G
Top: CATCGT
Bottom: CCCCCC

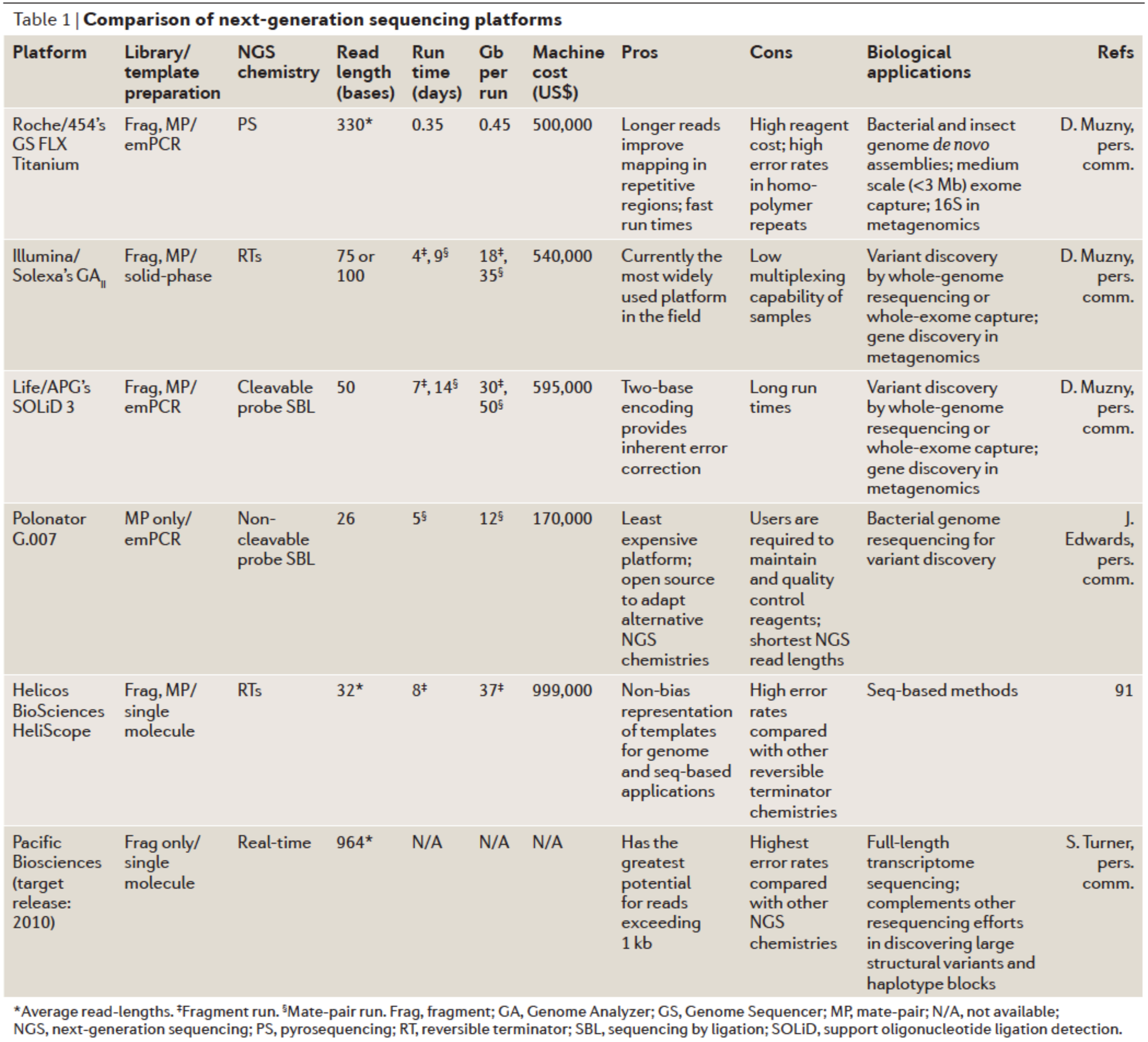


Table 1 | **Comparison of next-generation sequencing platforms**

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA$_{II}$	Frag, MP/ solid-phase	RTs	75 or 100	4‡, 9§	18‡, 35§	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7‡, 14§	30‡, 50§	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Polonator G.007	MP only/ emPCR	Non-cleavable probe SBL	26	5§	12§	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery	J. Edwards, pers. comm.
Helicos BioSciences HeliScope	Frag, MP/ single molecule	RTs	32*	8‡	37‡	999,000	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods	91
Pacific Biosciences (target release: 2010)	Frag only/ single molecule	Real-time	964*	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks	S. Turner, pers. comm.

*Average read-lengths. ‡Fragment run. §Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection.

(2009)

Challenges: Short Reads Alignment (or Mapping problem)

- Speed
  - How to map millions of short reads against a reference genome (*practicality*)
- Strategic
  - How to avoid mapping a read to multiple regions in the genome (*confidence*)

Short Reads Mapping Tools

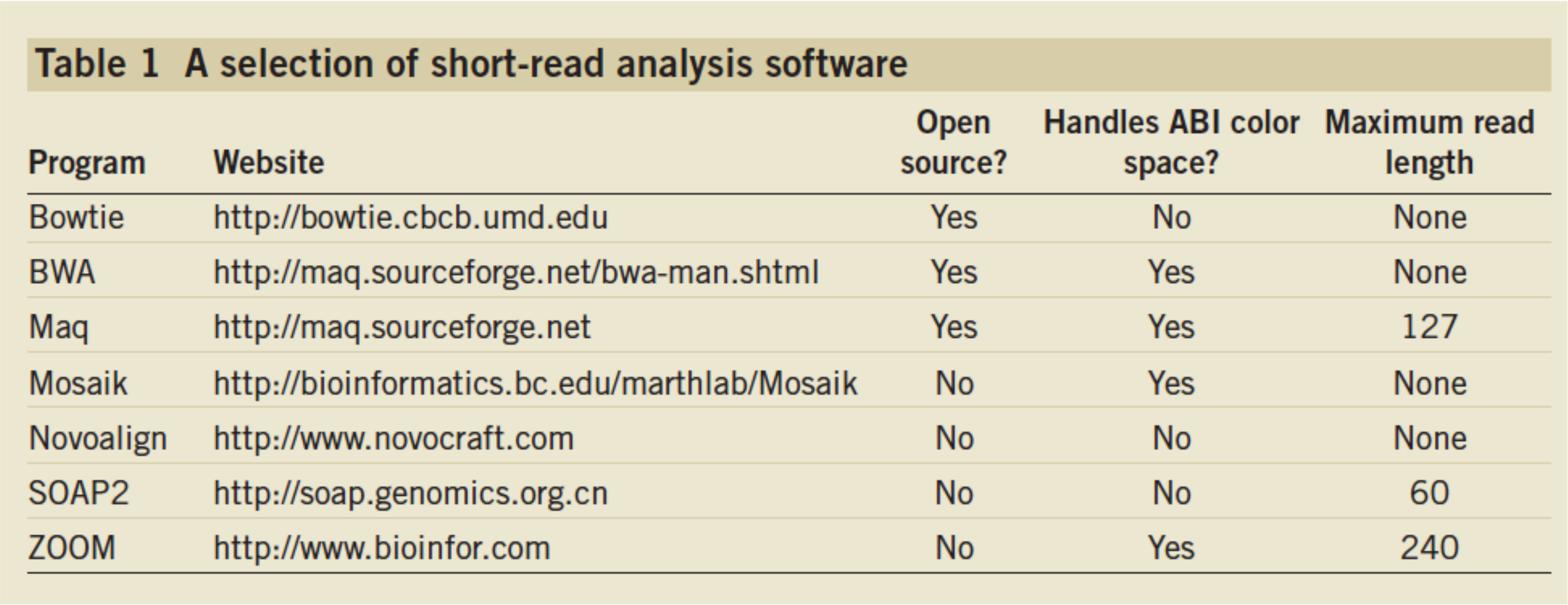


Table 1 A selection of short-read analysis software

Program	Website	Open source?	Handles ABI color space?	Maximum read length
Bowtie	http://bowtie.cbcb.umd.edu	Yes	No	None
BWA	http://maq.sourceforge.net/bwa-man.shtml	Yes	Yes	None
Maq	http://maq.sourceforge.net	Yes	Yes	127
Mosaik	http://bioinformatics.bc.edu/marthlab/Mosaik	No	Yes	None
Novoalign	http://www.novocraft.com	No	No	None
SOAP2	http://soap.genomics.org.cn	No	No	60
ZOOM	http://www.bioinfor.com	No	Yes	240

Two Strategies

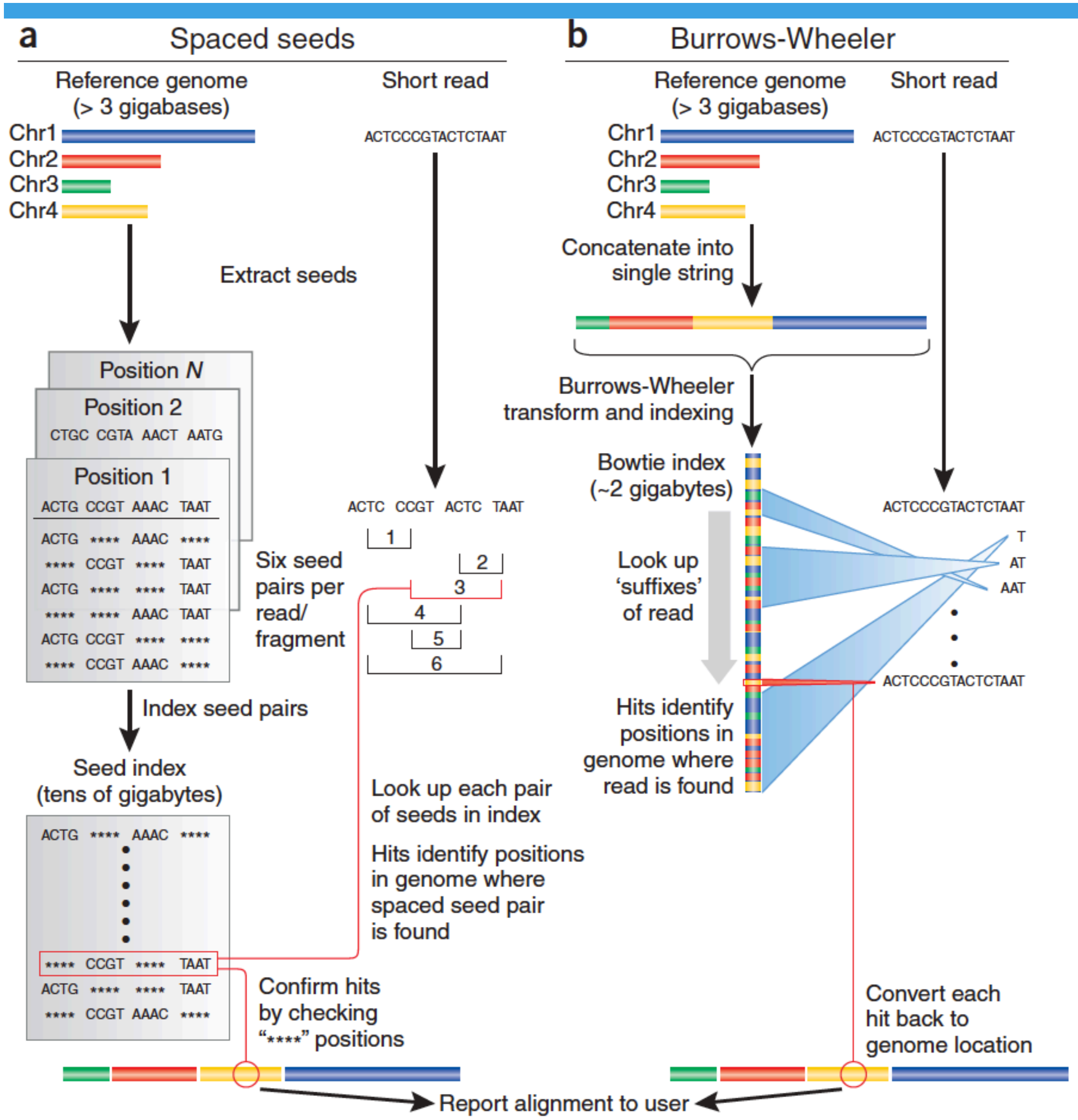
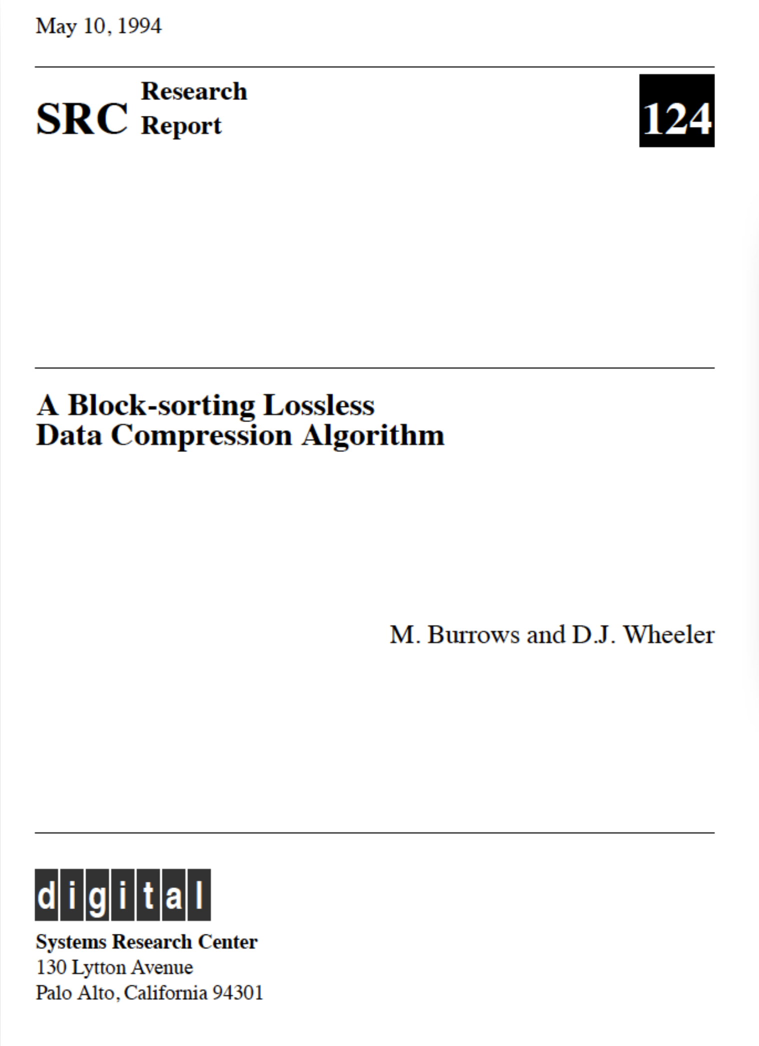


Figure 1 Two recent algorithmic approaches for aligning short (20–200-bp) sequencing reads. **(a)** Algorithms based on spaced-seed indexing, such as Maq, index the reads as follows: each position in the reference is cut into equal-sized pieces, called 'seeds' and these seeds are paired and stored in a lookup table. Each read is also cut up according to this scheme, and pairs of seeds are used as keys to look up matching positions in the reference. Because seed indices can be very large, some algorithms (including Maq) index the reads in batches and treat substrings of the reference as queries. **(b)** Algorithms based on the Burrows-Wheeler transform, such as Bowtie, store a memory-efficient representation of the reference genome. Reads are aligned character by character from right to left against the transformed string. With each new character, the algorithm updates an interval (indicated by blue 'beams') in the transformed string. When all characters in the read have been processed, alignments are represented by any positions within the interval. Burrows-Wheeler–based algorithms can run substantially faster than spaced seed approaches, primarily owing to the memory efficiency of the Burrows-Wheeler search. Chr., chromosome.

Burrows-Wheeler Algorithm



May 10, 1994

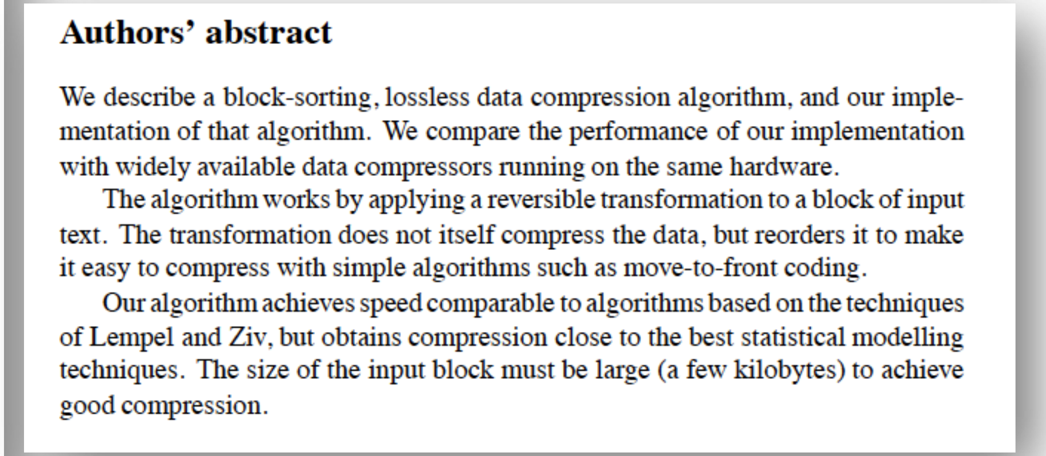
SRC Research Report 124

A Block-sorting Lossless Data Compression Algorithm

M. Burrows and D.J. Wheeler

digital

Systems Research Center
130 Lytton Avenue
Palo Alto, California 94301



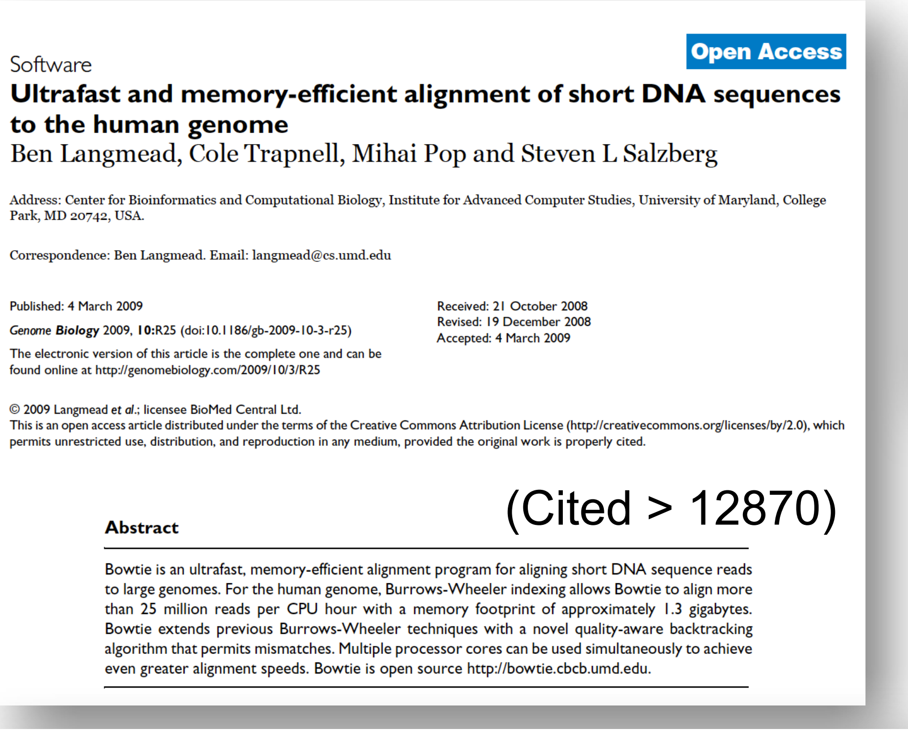
Authors' abstract

We describe a block-sorting, lossless data compression algorithm, and our implementation of that algorithm. We compare the performance of our implementation with widely available data compressors running on the same hardware.

The algorithm works by applying a reversible transformation to a block of input text. The transformation does not itself compress the data, but reorders it to make it easy to compress with simple algorithms such as move-to-front coding.

Our algorithm achieves speed comparable to algorithms based on the techniques of Lempel and Ziv, but obtains compression close to the best statistical modelling techniques. The size of the input block must be large (a few kilobytes) to achieve good compression.

Bowtie and BWA



Software

Open Access

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg

Address: Center for Bioinformatics and Computational Biology, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA.

Correspondence: Ben Langmead. Email: langmead@cs.umd.edu

Published: 4 March 2009

Genome Biology 2009, **10:**R25 (doi:10.1186/gb-2009-10-3-r25)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/3/R25

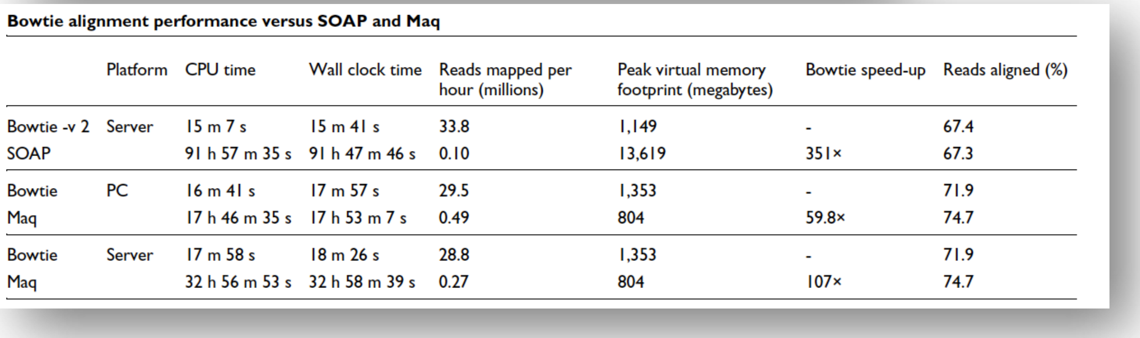
Received: 21 October 2008
Revised: 19 December 2008
Accepted: 4 March 2009

© 2009 Langmead *et al.*; licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

(Cited > 12870)

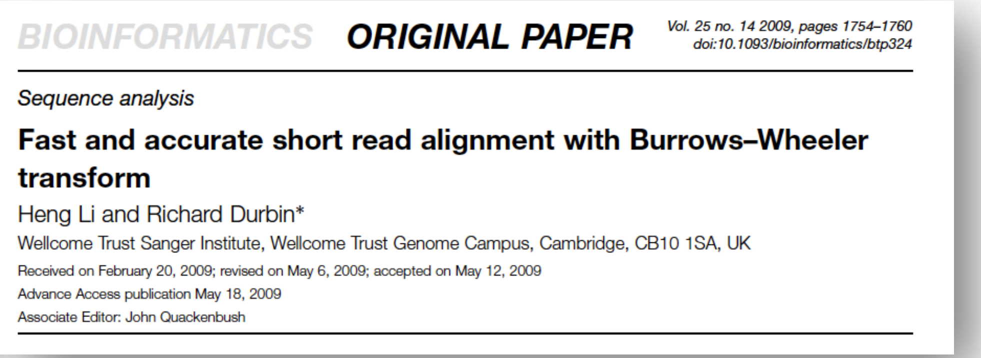
Abstract

Bowtie is an ultrafast, memory-efficient alignment program for aligning short DNA sequence reads to large genomes. For the human genome, Burrows-Wheeler indexing allows Bowtie to align more than 25 million reads per CPU hour with a memory footprint of approximately 1.3 gigabytes. Bowtie extends previous Burrows-Wheeler techniques with a novel quality-aware backtracking algorithm that permits mismatches. Multiple processor cores can be used simultaneously to achieve even greater alignment speeds. Bowtie is open source http://bowtie.cbcb.umd.edu.



Bowtie alignment performance versus SOAP and Maq

	Platform	CPU time	Wall clock time	Reads mapped per hour (millions)	Peak virtual memory footprint (megabytes)	Bowtie speed-up	Reads aligned (%)
Bowtie -v 2	Server	15 m 7 s	15 m 41 s	33.8	1,149	-	67.4
SOAP		91 h 57 m 35 s	91 h 47 m 46 s	0.10	13,619	351×	67.3
Bowtie	PC	16 m 41 s	17 m 57 s	29.5	1,353	-	71.9
Maq		17 h 46 m 35 s	17 h 53 m 7 s	0.49	804	59.8×	74.7
Bowtie	Server	17 m 58 s	18 m 26 s	28.8	1,353	-	71.9
Maq		32 h 56 m 53 s	32 h 58 m 39 s	0.27	804	107×	74.7



BIOINFORMATICS **ORIGINAL PAPER**

Vol. 25 no. 14 2009, pages 1754–1760
doi:10.1093/bioinformatics/btp324

Sequence analysis

Fast and accurate short read alignment with Burrows–Wheeler transform

Heng Li and Richard Durbin*

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK

Received on February 20, 2009; revised on May 6, 2009; accepted on May 12, 2009

Advance Access publication May 18, 2009

Associate Editor: John Quackenbush

(Cited > 16200)

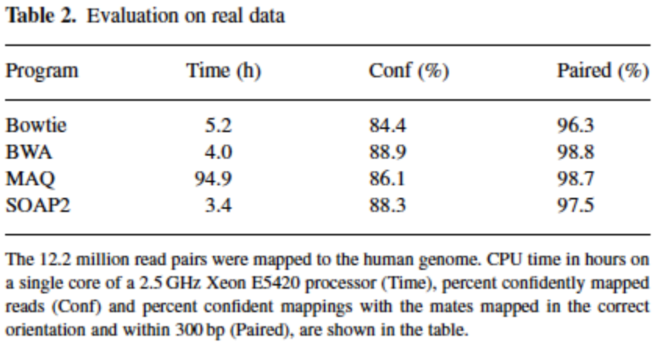


Table 2. Evaluation on real data

Program	Time (h)	Conf (%)	Paired (%)
Bowtie	5.2	84.4	96.3
BWA	4.0	88.9	98.8
MAQ	94.9	86.1	98.7
SOAP2	3.4	88.3	97.5

The 12.2 million read pairs were mapped to the human genome. CPU time in hours on a single core of a 2.5 GHz Xeon E5420 processor (Time), percent confidently mapped reads (Conf) and percent confident mappings with the mates mapped in the correct orientation and within 300 bp (Paired), are shown in the table.

Burrows-Wheeler Transform (BWT)

Text transform that is useful for compression & search.

BANANA

BANANA$
ANANA$B
NANA$BA
ANA$BAN
NA$BANA
A$BANAN
$BANANA

→

$BANANA
A$BANAN
ANA$BAN
ANANA$B
BANANA$
NA$BANA
NANA$BA

BWT(BANANA) =

ANNB$AA

Tends to put runs of the same character together.

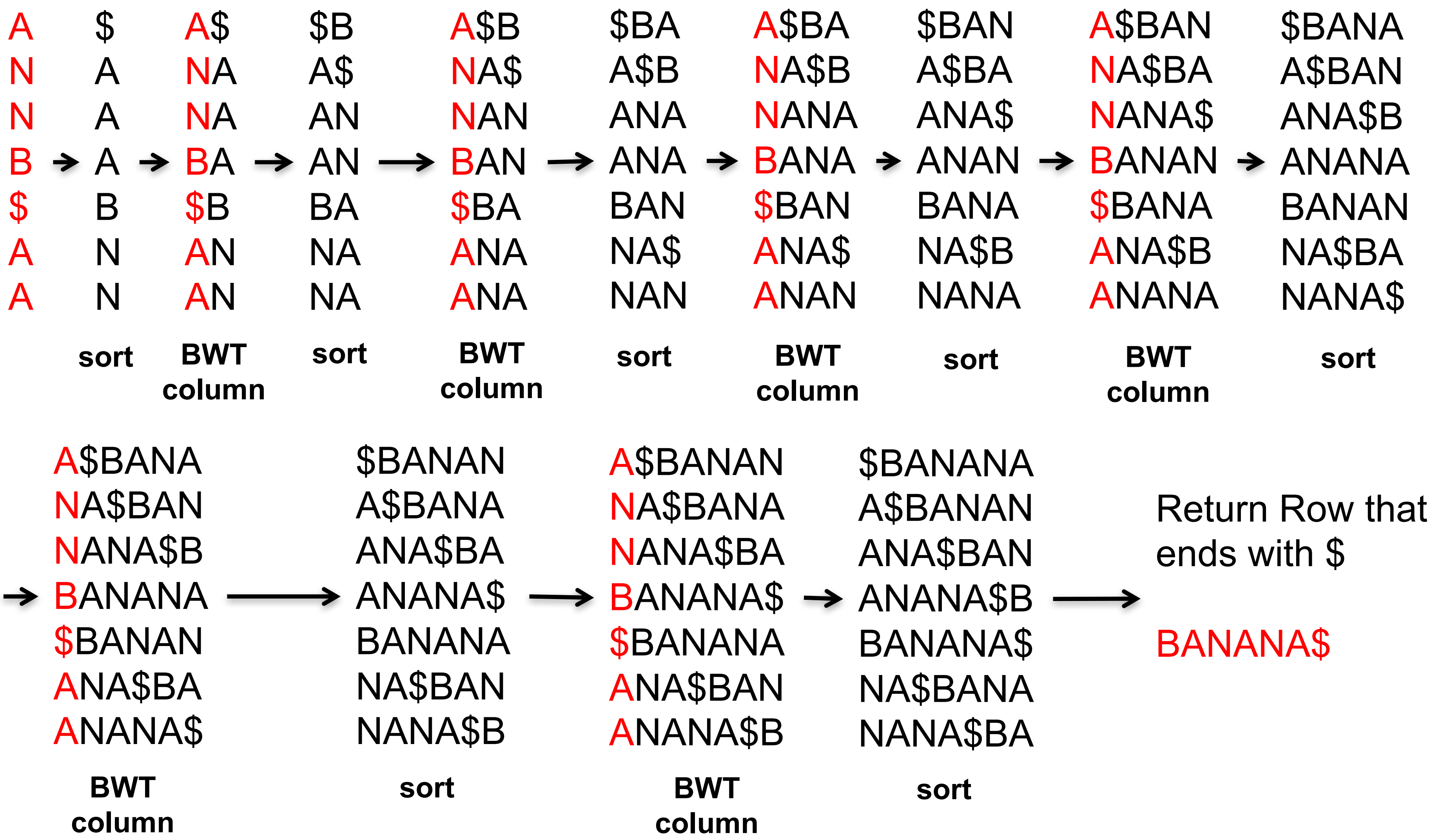
Makes compression work well.

“bzip” is based on this.

(Adapted from CMSC 858s: Computational Genomics slides by Carl Kingsford)

Burrows-Wheeler Transform (BWT)

Recovering ANNB$AA



A	$	A$	$B	A$B	$BA	A$BA	$BAN	A$BAN	$BANA
N	A	NA	A$	NA$	A$B	NA$B	A$BA	NA$BA	A$BAN
N	A	NA	AN	NAN	ANA	NANA	ANA$	NANA$	ANA$B
B	A	BA	AN	BAN	ANA	BANA	ANAN	BANAN	ANANA
$	B	$B	BA	$BA	BAN	$BAN	BANA	$BANA	BANAN
A	N	AN	NA	ANA	NA$	ANA$	NA$B	ANA$B	NA$BA
A	N	AN	NA	ANA	NAN	ANAN	NANA	ANANA	NANA$
	sort	**BWT column**	**sort**	**BWT column**	**sort**	**BWT column**	**sort**	**BWT column**	**sort**

A$BANA	$BANAN	A$BANAN	$BANANA
NA$BAN	A$BANA	NA$BANA	A$BANAN
NANA$B	ANA$BA	NANA$BA	ANA$BAN
BANANA	ANANA$	BANANA$	ANANA$B
$BANAN	BANANA	$BANANA	BANANA$
ANA$BA	NA$BAN	ANA$BAN	NA$BANA
ANANA$	NANA$B	ANANA$B	NANA$BA
BWT column	**sort**	**BWT column**	**sort**

Return Row that ends with $

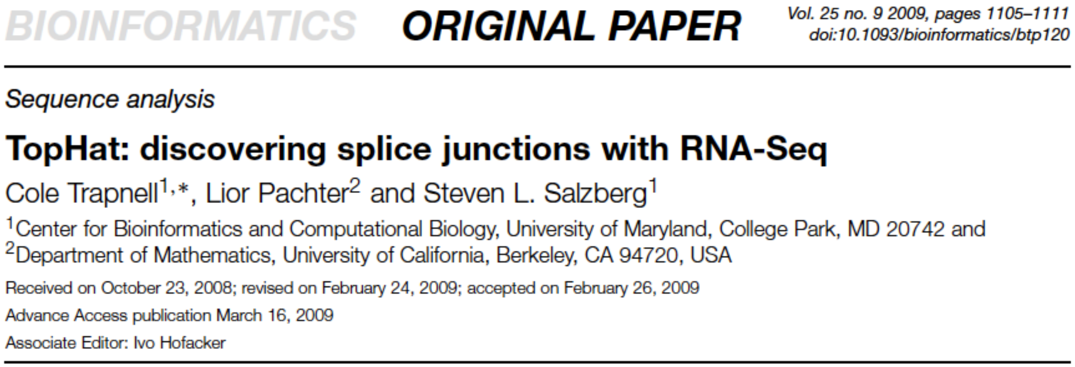
BANANA$

BWT Algorithm

- BWT useful for searching and compression.
- BWT is invertible: given the BWT of a string, the string can be reconstructed.
- BWT is computable in O(n) time.
- Even after compression, can search string quickly.

(Adapted from CMSC 858s: Computational Genomics slides by Carl Kingsford)

TopHat



BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 9 2009, pages 1105–1111
doi:10.1093/bioinformatics/btp120

Sequence analysis

TopHat: discovering splice junctions with RNA-Seq

Cole Trapnell[1],*, Lior Pachter[2] and Steven L. Salzberg[1]

[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742 and
[2]Department of Mathematics, University of California, Berkeley, CA 94720, USA

Received on October 23, 2008; revised on February 24, 2009; accepted on February 26, 2009
Advance Access publication March 16, 2009
Associate Editor: Ivo Hofacker

(Cited > 7900)

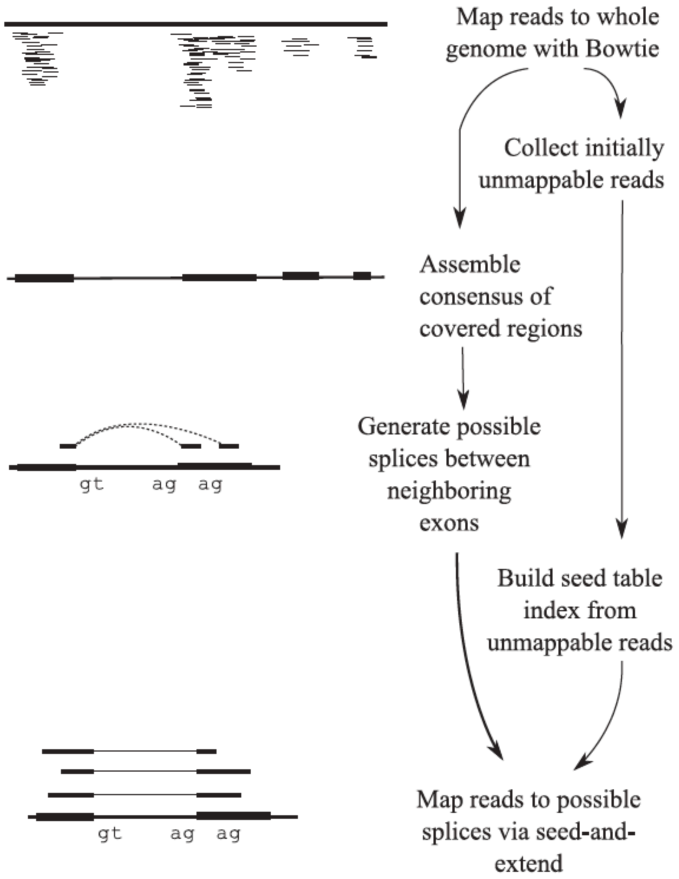

Map reads to whole genome with Bowtie
Collect initially unmappable reads
Assemble consensus of covered regions
Generate possible splices between neighboring exons
gt
ag
ag
Build seed table index from unmappable reads
Map reads to possible splices via seed-and-extend
gt
ag
ag

Fig. 1. The TopHat pipeline. RNA-Seq reads are mapped against the whole reference genome, and those reads that do not map are set aside. An initial consensus of mapped regions is computed by Maq. Sequences flanking potential donor/acceptor splice sites within neighboring regions are joined to form potential splice junctions. The IUM reads are indexed and aligned to these splice junction sequences.

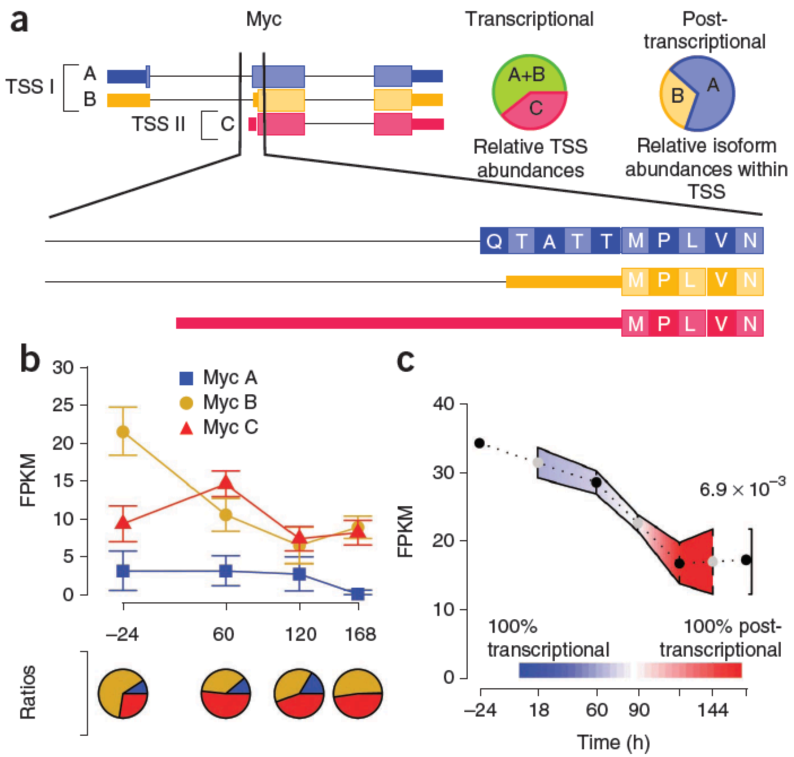
Cufflinks

Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

Cole Trapnell[1–3], Brian A Williams[4], Geo Pertea[2], Ali Mortazavi[4], Gordon Kwan[4], Marijke J van Baren[5], Steven L Salzberg[1,2], Barbara J Wold[4] & Lior Pachter[3,6,7]

Nature Biotechnology 2010. 28 (5), 511-515 (Cited > 7200)

High-throughput mRNA sequencing (RNA-Seq) promises simultaneous transcript discovery and abundance estimation[1–3]. However, this would require algorithms that are not restricted by prior gene annotations and that account for alternative transcription and splicing. Here we introduce such algorithms in an open-source software program called Cufflinks. To test Cufflinks, we sequenced and analyzed >430 million paired 75-bp RNA-Seq reads from a mouse myoblast cell line over a differentiation time series. We detected 13,692 known transcripts and 3,724 previously unannotated ones, 62% of which are supported by independent expression data or by homologous genes in other species. Over the time series, 330 genes showed complete switches in the dominant transcription start site (TSS) or splice isoform, and we observed more subtle shifts in 1,304 other genes. These results suggest that Cufflinks can illuminate the substantial regulatory flexibility and complexity in even this well-studied model of muscle development and that it can improve transcriptome-based genome annotation.



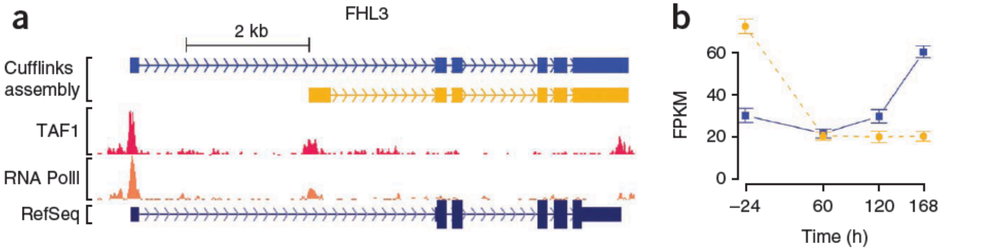
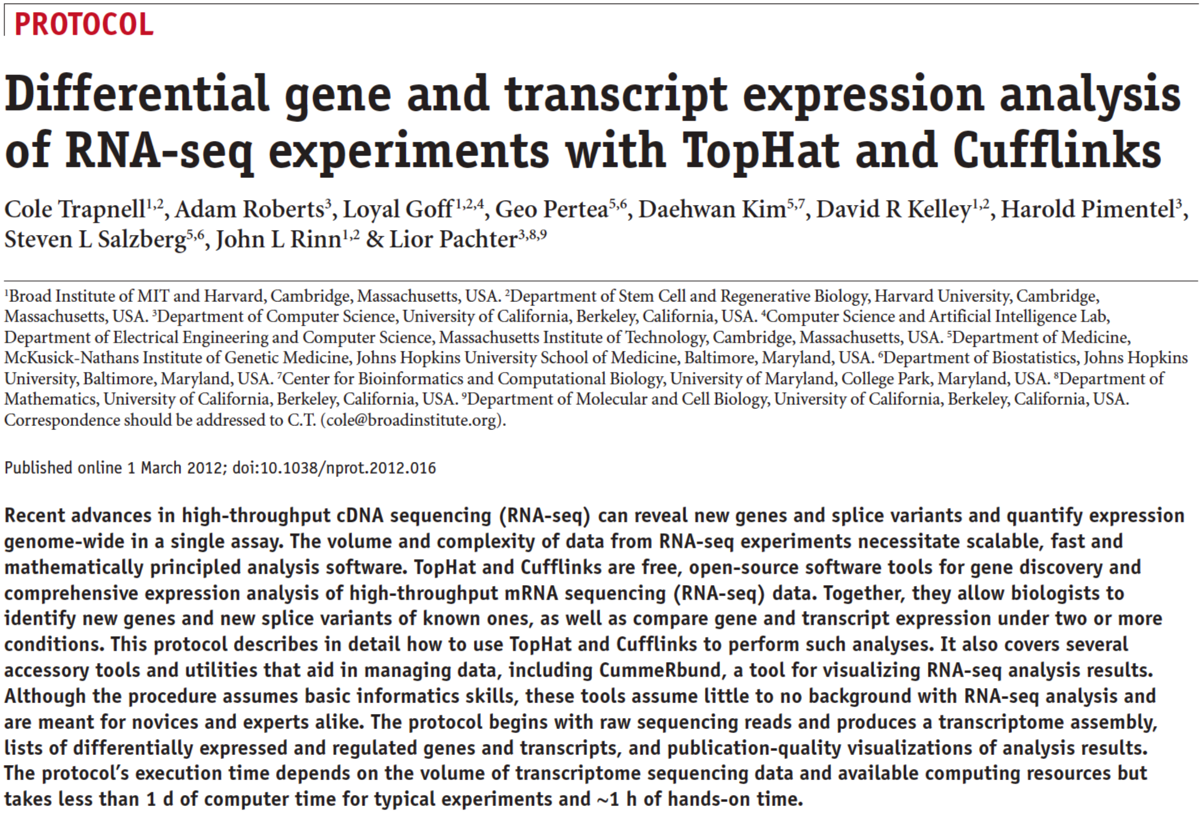


Figure 3 Excluding isoforms discovered by Cufflinks from the transcript abundance estimation affects the abundance estimates of known isoforms, in some cases by orders of magnitude. FHL3 inhibits myogenesis by binding MyoD and attenuating its transcriptional activity. (**a**) The C2C12 transcriptome contains a novel isoform that is dominant during proliferation. The new TSS for FHL3 is supported by proximal TAF1 and RNA polymerase II ChIP-Seq peaks. (**b**) The known isoform (solid line) is preferred at time points following differentiation.

Bowtie/TopHat/Cufflinks Workflow



PROTOCOL

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

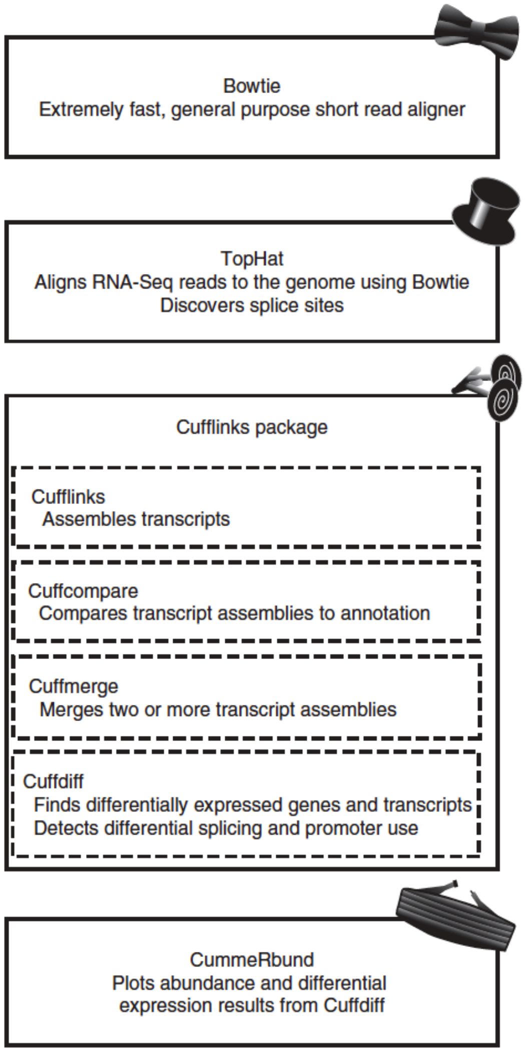
Cole Trapnell[1,2], Adam Roberts[3], Loyal Goff[1,2,4], Geo Pertea[5,6], Daehwan Kim[5,7], David R Kelley[1,2], Harold Pimentel[3], Steven L Salzberg[5,6], John L Rinn[1,2] & Lior Pachter[3,8,9]

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. [3]Department of Computer Science, University of California, Berkeley, California, USA. [4]Computer Science and Artificial Intelligence Lab, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Department of Medicine, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [6]Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA. [7]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. [8]Department of Mathematics, University of California, Berkeley, California, USA. [9]Department of Molecular and Cell Biology, University of California, Berkeley, California, USA. Correspondence should be addressed to C.T. (cole@broadinstitute.org).

Published online 1 March 2012; doi:10.1038/nprot.2012.016

Recent advances in high-throughput cDNA sequencing (RNA-seq) can reveal new genes and splice variants and quantify expression genome-wide in a single assay. The volume and complexity of data from RNA-seq experiments necessitate scalable, fast and mathematically principled analysis software. TopHat and Cufflinks are free, open-source software tools for gene discovery and comprehensive expression analysis of high-throughput mRNA sequencing (RNA-seq) data. Together, they allow biologists to identify new genes and new splice variants of known ones, as well as compare gene and transcript expression under two or more conditions. This protocol describes in detail how to use TopHat and Cufflinks to perform such analyses. It also covers several accessory tools and utilities that aid in managing data, including CummeRbund, a tool for visualizing RNA-seq analysis results. Although the procedure assumes basic informatics skills, these tools assume little to no background with RNA-seq analysis and are meant for novices and experts alike. The protocol begins with raw sequencing reads and produces a transcriptome assembly, lists of differentially expressed and regulated genes and transcripts, and publication-quality visualizations of analysis results. The protocol's execution time depends on the volume of transcriptome sequencing data and available computing resources but takes less than 1 d of computer time for typical experiments and ~1 h of hands-on time.

(Cited > 5600)



Bowtie
Extremely fast, general purpose short read aligner

TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites

Cufflinks package

Cufflinks
Assembles transcripts

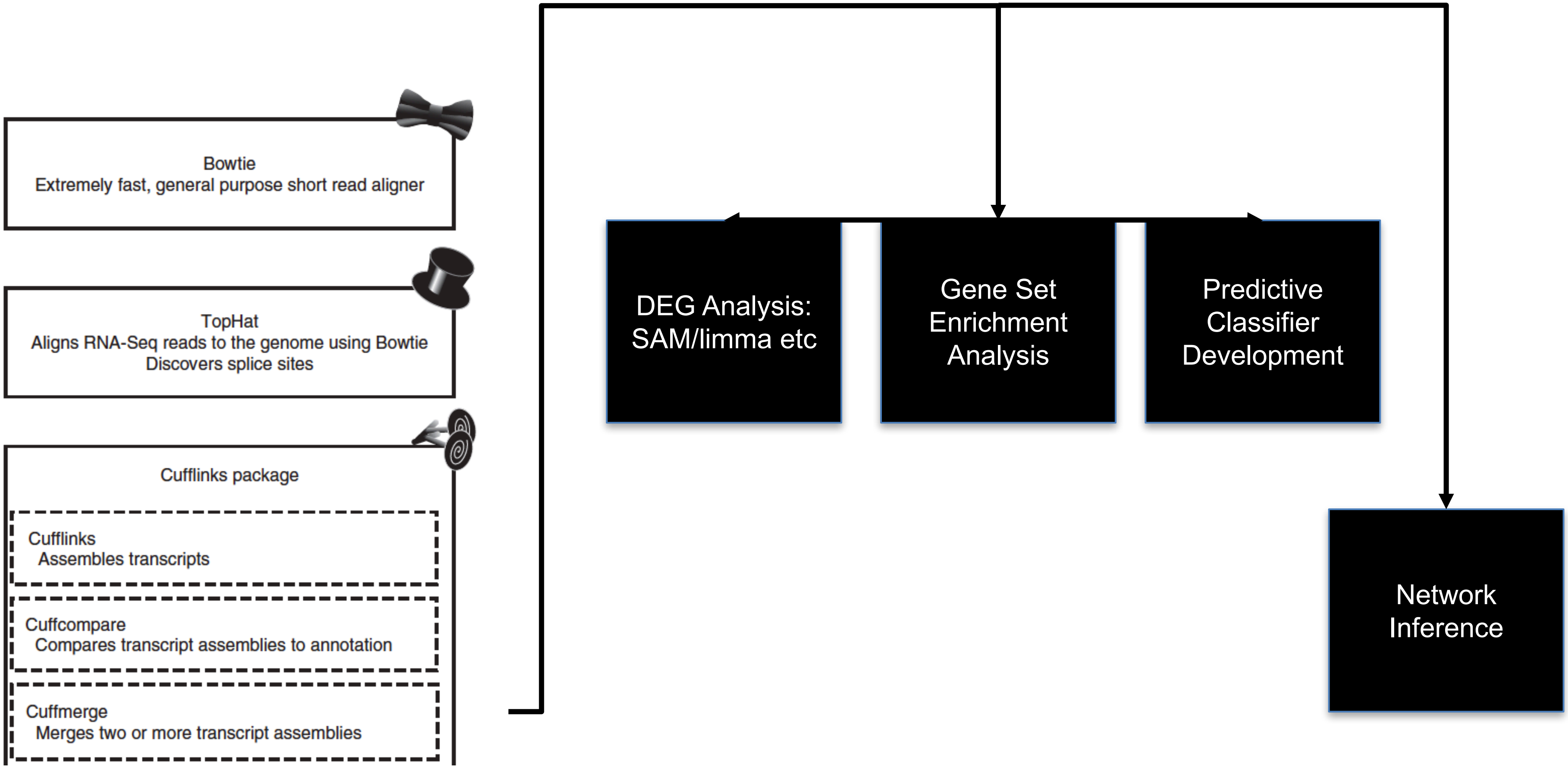
Cuffcompare
Compares transcript assemblies to annotation

Cuffmerge
Merges two or more transcript assemblies

Cuffdiff
Finds differentially expressed genes and transcripts
Detects differential splicing and promoter use

CummeRbund
Plots abundance and differential expression results from Cuffdiff

Tuxedo Workflow for RNA-seq Analysis



Bowtie
Extremely fast, general purpose short read aligner

TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites

Cufflinks package

Cufflinks
Assembles transcripts

Cuffcompare
Compares transcript assemblies to annotation

Cuffmerge
Merges two or more transcript assemblies

DEG Analysis: SAM/limma etc

Gene Set Enrichment Analysis

Predictive Classifier Development

Network Inference

Transcript expression is quantified in FPKM (fragment per kilobase of exon per million mapped reads)

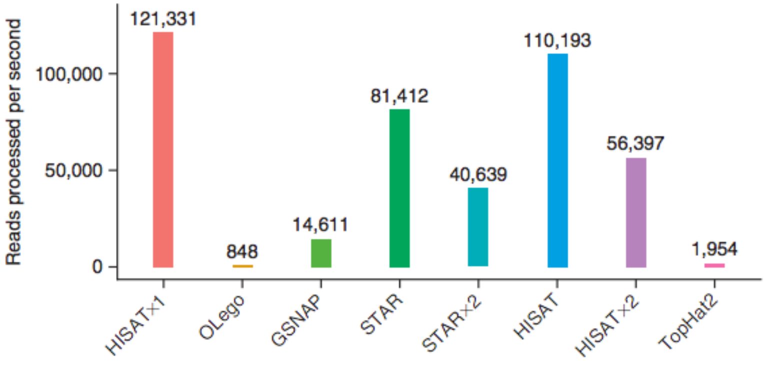
HISAT

HISAT: a fast spliced aligner with low memory requirements

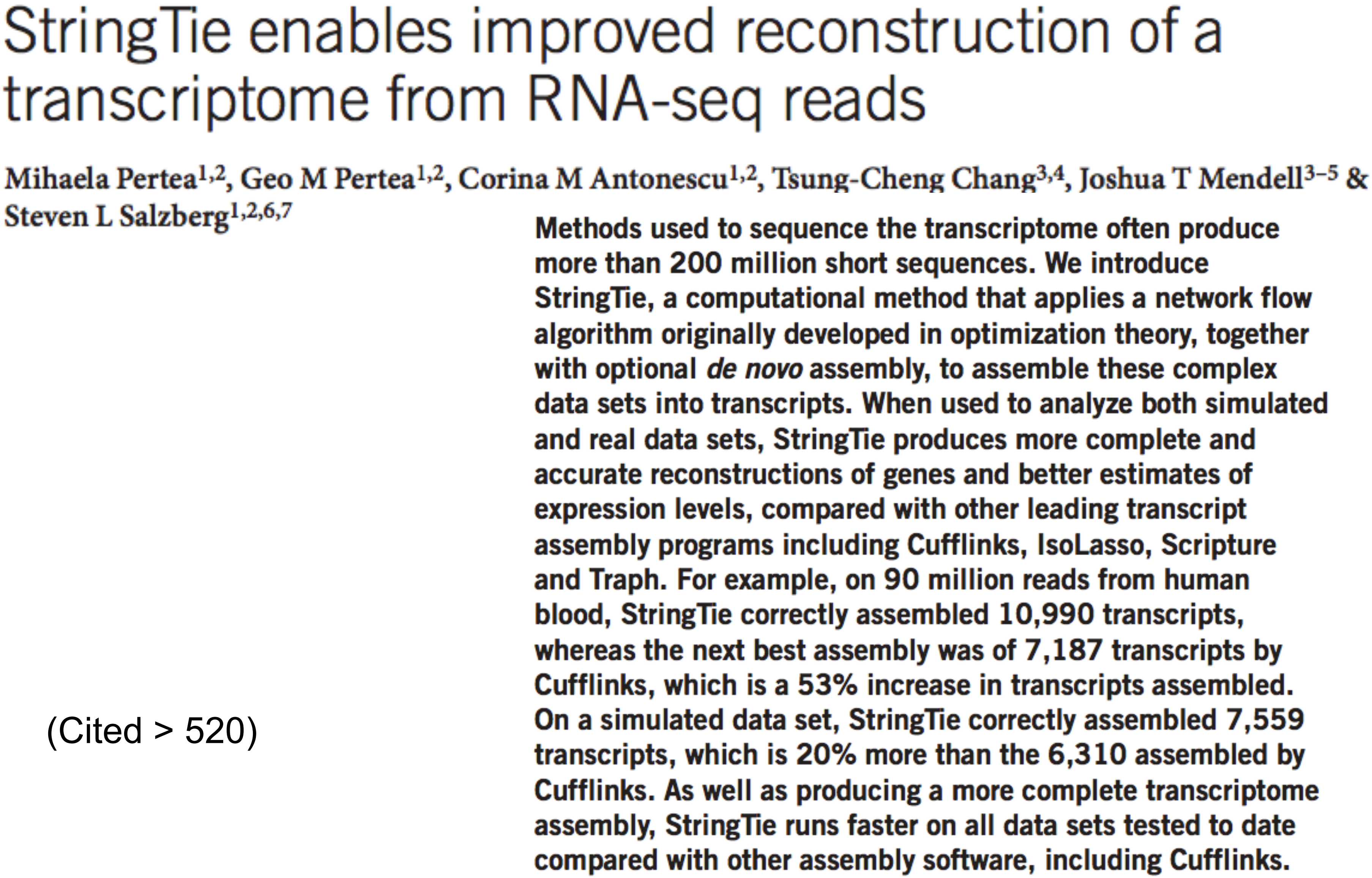
(Cited > 1150)

Daehwan Kim[1,2], Ben Langmead[1–3] & Steven L Salzberg[1–3]

HISAT (hierarchical indexing for spliced alignment of transcripts) is a highly efficient system for aligning reads from RNA sequencing experiments. HISAT uses an indexing scheme based on the Burrows-Wheeler transform and the Ferragina-Manzini (FM) index, employing two types of indexes for alignment: a whole-genome FM index to anchor each alignment and numerous local FM indexes for very rapid extensions of these alignments. HISAT's hierarchical index for the human genome contains 48,000 local FM indexes, each representing a genomic region of ~64,000 bp. Tests on real and simulated data sets showed that HISAT is the fastest system currently available, with equal or better accuracy than any other method. Despite its large number of indexes, HISAT requires only 4.3 gigabytes of memory. HISAT supports genomes of any size, including those larger than 4 billion bases.



StringTie



StringTie enables improved reconstruction of a transcriptome from RNA-seq reads

Mihaela Pertea[1,2], Geo M Pertea[1,2], Corina M Antonescu[1,2], Tsung-Cheng Chang[3,4], Joshua T Mendell[3–5] & Steven L Salzberg[1,2,6,7]

Methods used to sequence the transcriptome often produce more than 200 million short sequences. We introduce StringTie, a computational method that applies a network flow algorithm originally developed in optimization theory, together with optional *de novo* assembly, to assemble these complex data sets into transcripts. When used to analyze both simulated and real data sets, StringTie produces more complete and accurate reconstructions of genes and better estimates of expression levels, compared with other leading transcript assembly programs including Cufflinks, IsoLasso, Scripture and Traph. For example, on 90 million reads from human blood, StringTie correctly assembled 10,990 transcripts, whereas the next best assembly was of 7,187 transcripts by Cufflinks, which is a 53% increase in transcripts assembled. On a simulated data set, StringTie correctly assembled 7,559 transcripts, which is 20% more than the 6,310 assembled by Cufflinks. As well as producing a more complete transcriptome assembly, StringTie runs faster on all data sets tested to date compared with other assembly software, including Cufflinks.

(Cited > 520)

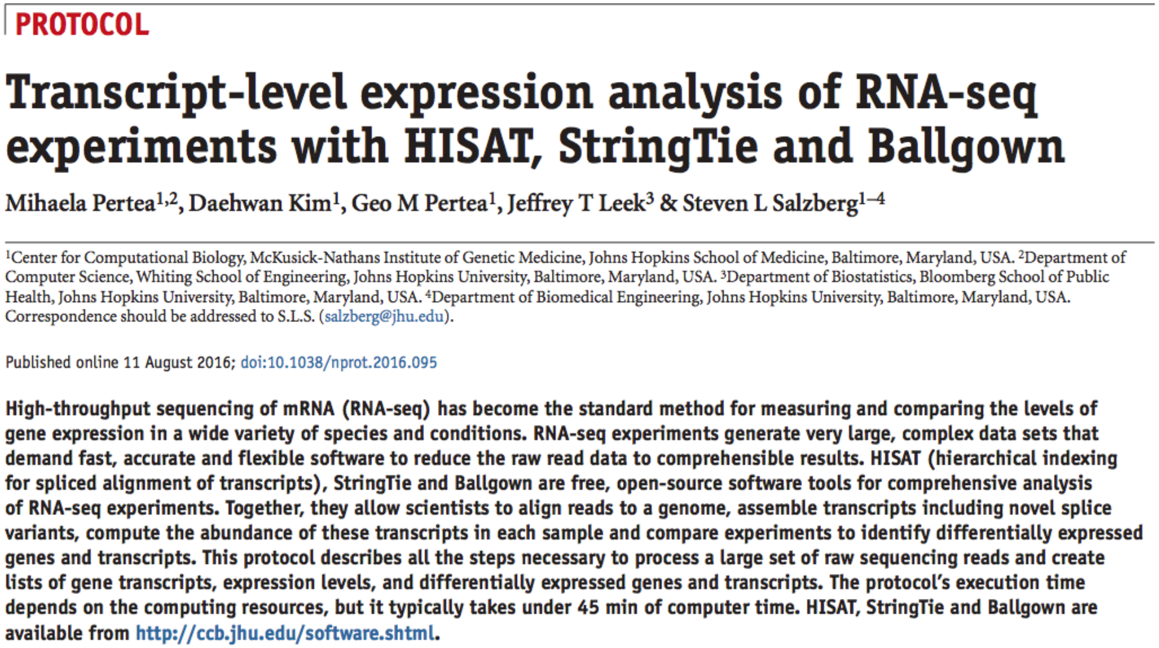
Ballgown

Ballgown bridges the gap between transcriptome assembly and expression analysis

(Cited > 90)

Ballgown can function as a bridge between upstream assembly tools, such as Cufflinks, and downstream statistical modeling tools in Bioconductor. The Ballgown suite includes functions for interactive exploration of the transcriptome assembly, visualization of transcript structures and feature-specific abundances for each locus and post hoc annotation of assembled features to annotated features. Direct availability of feature-by-sample expression tables makes it easy to apply alternative differential expression tests or to evaluate other statistical properties of the assembly, such as dispersion of expression values across replicates or genes. The Tablemaker preprocessor writes the tables directly to disk, and they can be loaded into R with a single function call. The Ballgown and Tablemaker software packages are available from Bioconductor and GitHub

RNA-seq: Adapted Bowtie/TopHat/Cufflinks Workflow



PROTOCOL

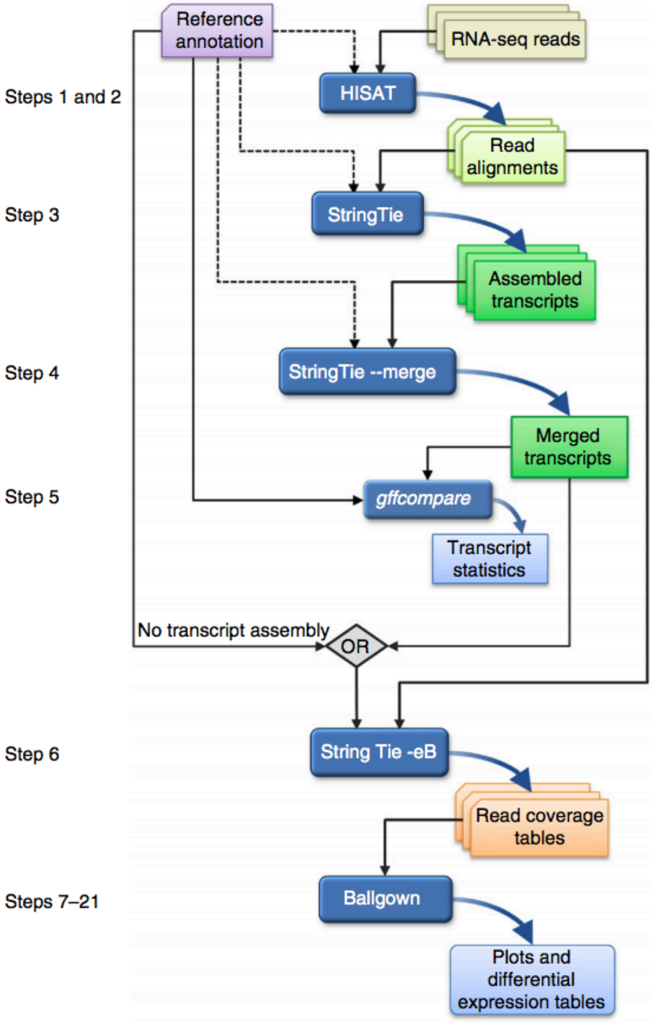
Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

Mihaela Pertea[1,2], Daehwan Kim[1], Geo M Pertea[1], Jeffrey T Leek[3] & Steven L Salzberg[1–4]

[1]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA. [2]Department of Computer Science, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland, USA. [3]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA. [4]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA. Correspondence should be addressed to S.L.S. (salzberg@jhu.edu).

Published online 11 August 2016; doi:10.1038/nprot.2016.095

High-throughput sequencing of mRNA (RNA-seq) has become the standard method for measuring and comparing the levels of gene expression in a wide variety of species and conditions. RNA-seq experiments generate very large, complex data sets that demand fast, accurate and flexible software to reduce the raw read data to comprehensible results. HISAT (hierarchical indexing for spliced alignment of transcripts), StringTie and Ballgown are free, open-source software tools for comprehensive analysis of RNA-seq experiments. Together, they allow scientists to align reads to a genome, assemble transcripts including novel splice variants, compute the abundance of these transcripts in each sample and compare experiments to identify differentially expressed genes and transcripts. This protocol describes all the steps necessary to process a large set of raw sequencing reads and create lists of gene transcripts, expression levels, and differentially expressed genes and transcripts. The protocol's execution time depends on the computing resources, but it typically takes under 45 min of computer time. HISAT, StringTie and Ballgown are available from http://ccb.jhu.edu/software.shtml.



(Cited > 250)

Applications of NGS

- Whole Genome Sequencing
- Exome Sequencing
- Genetic Variations
- Transcriptome variations (gene expression, isoforms, gene fusions)
- Gene regulations (TF binding sites, PolII binding patterns, miRNA-mRNA interactions etc)
- Epigenetic (nucleosome positioning, genome-wide methylation patterns etc)
- Other functional genomics screens (shRNAs, siRNAs, etc)

RNA-seq

- Using NGS to sequence transcriptome (complete set of transcripts in a cell)
- **Goals:**
  - Discover full set of transcripts

    large & small RNA, coding and non-coding, novel transcripts, gene-fusion transcripts, sense and antisense transcripts, alternative splicing
  - Compare experimental conditions

    Differential expression (gene, isoform, splicing)

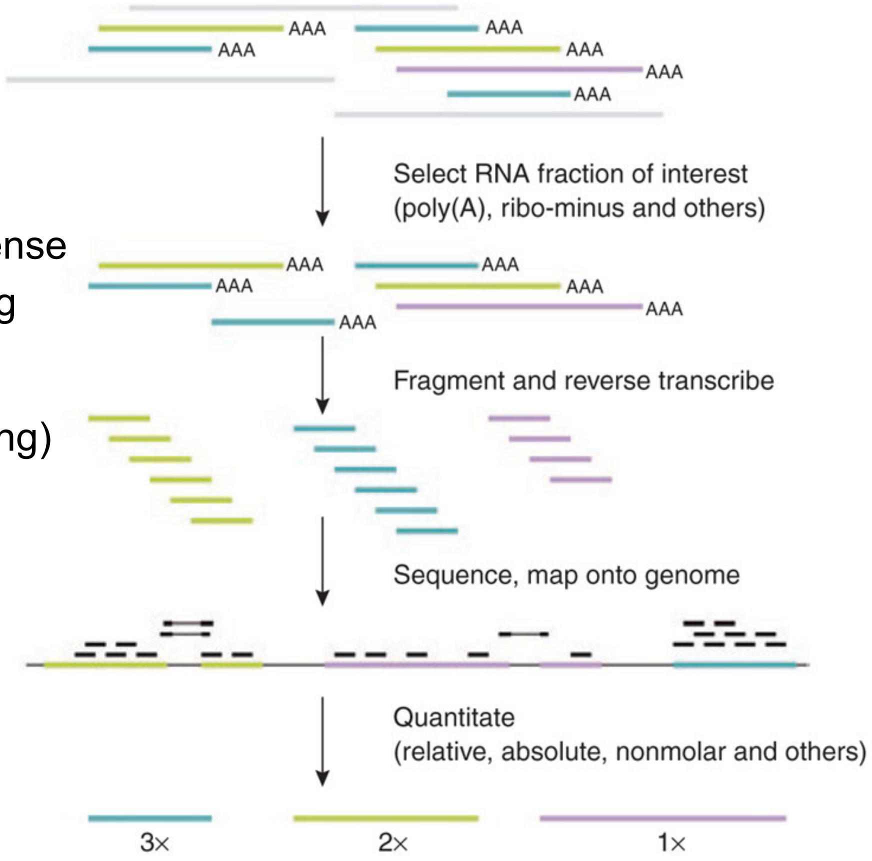

AAA
Select RNA fraction of interest
(poly(A), ribo-minus and others)
Fragment and reverse transcribe
Sequence, map onto genome
Quantitate
(relative, absolute, nonmolar and others)
3×
2×
1×

Figure 5 Pepke et al. (2009) *Nature Methods*

Oshlack et al., (2010) *Genome Biology* 11:220
Ozsolak & Milos (2011) *Nature Reviews Genetics* 12:87-98
Garber et al., (2011) *Nature Methods* 8:

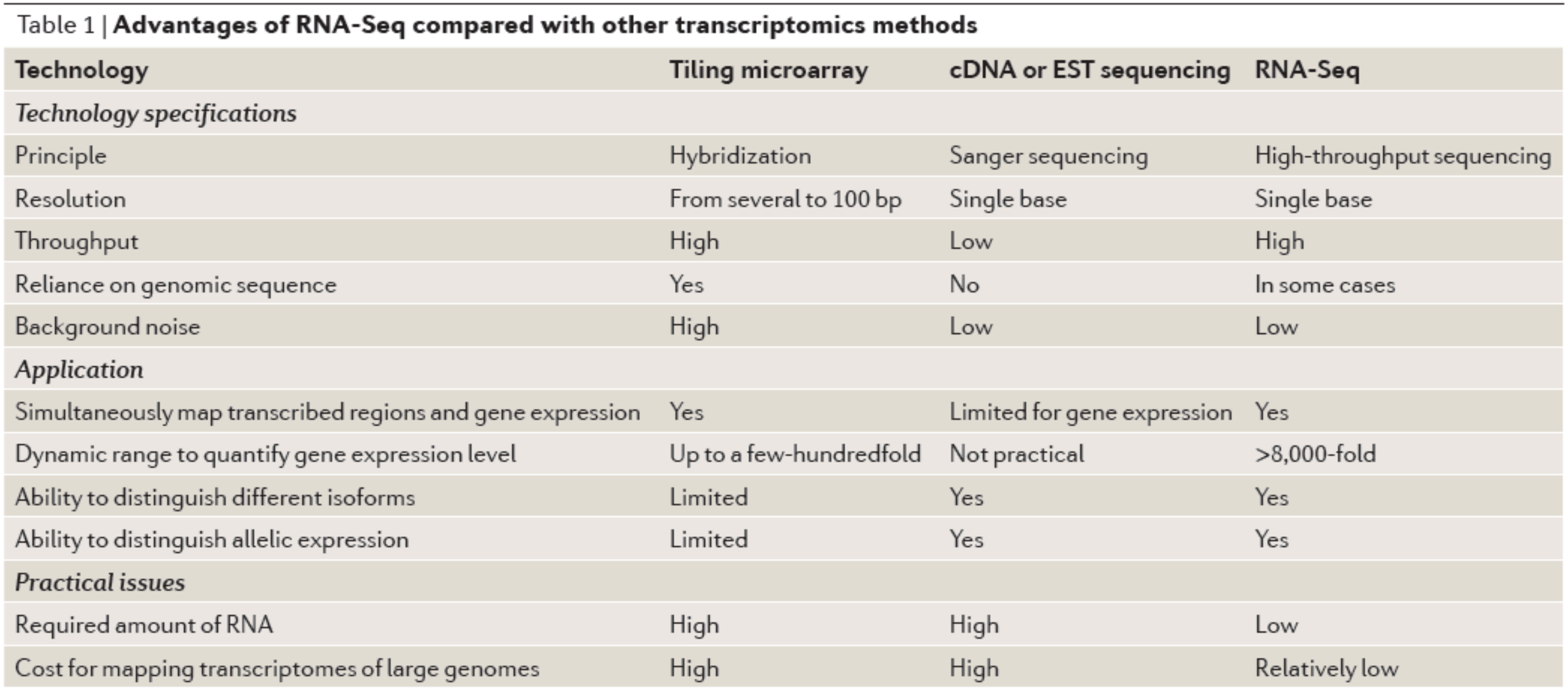


Table 1 | **Advantages of RNA-Seq compared with other transcriptomics methods**

Technology	**Tiling microarray**	**cDNA or EST sequencing**	**RNA-Seq**
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

Comparison of RNA-seq and Microarray

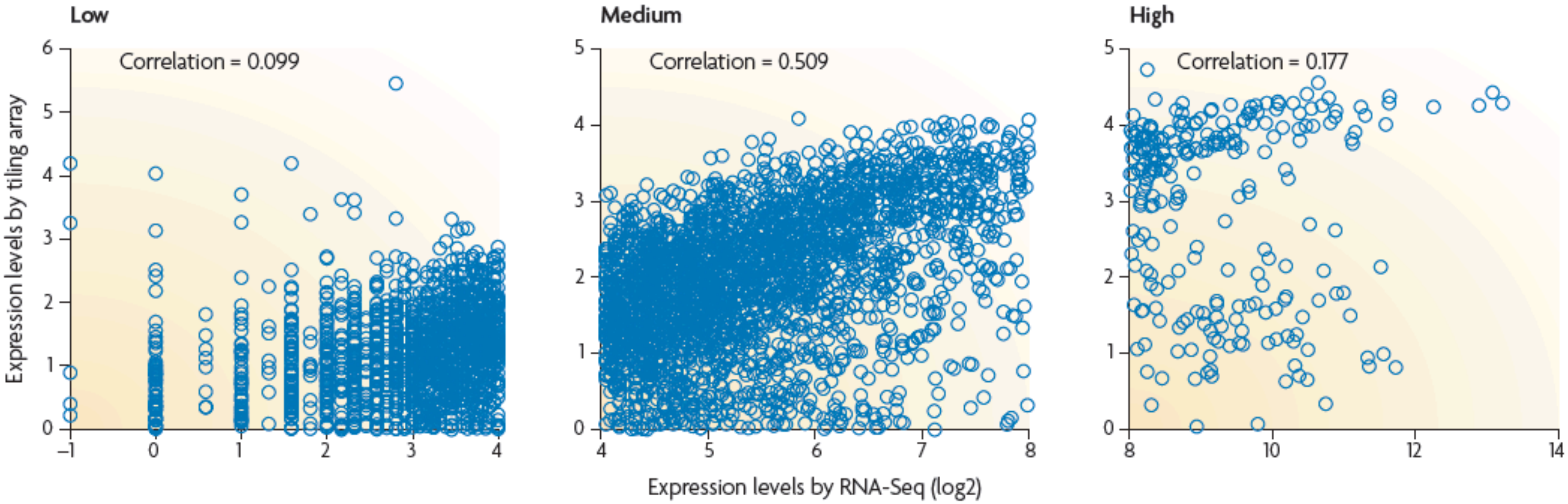
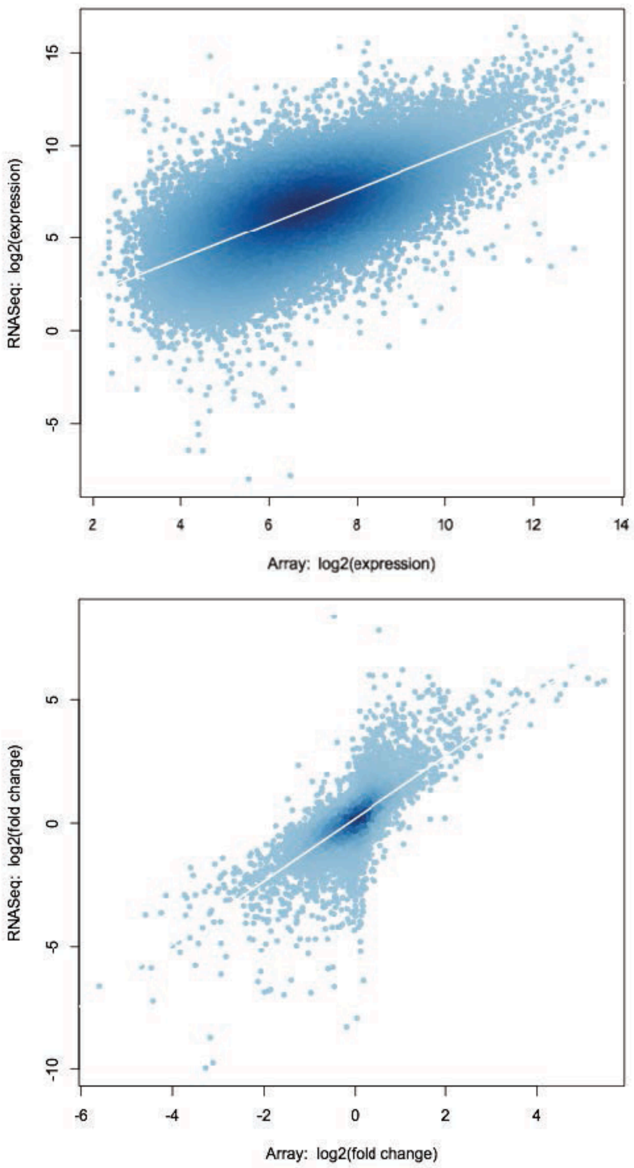


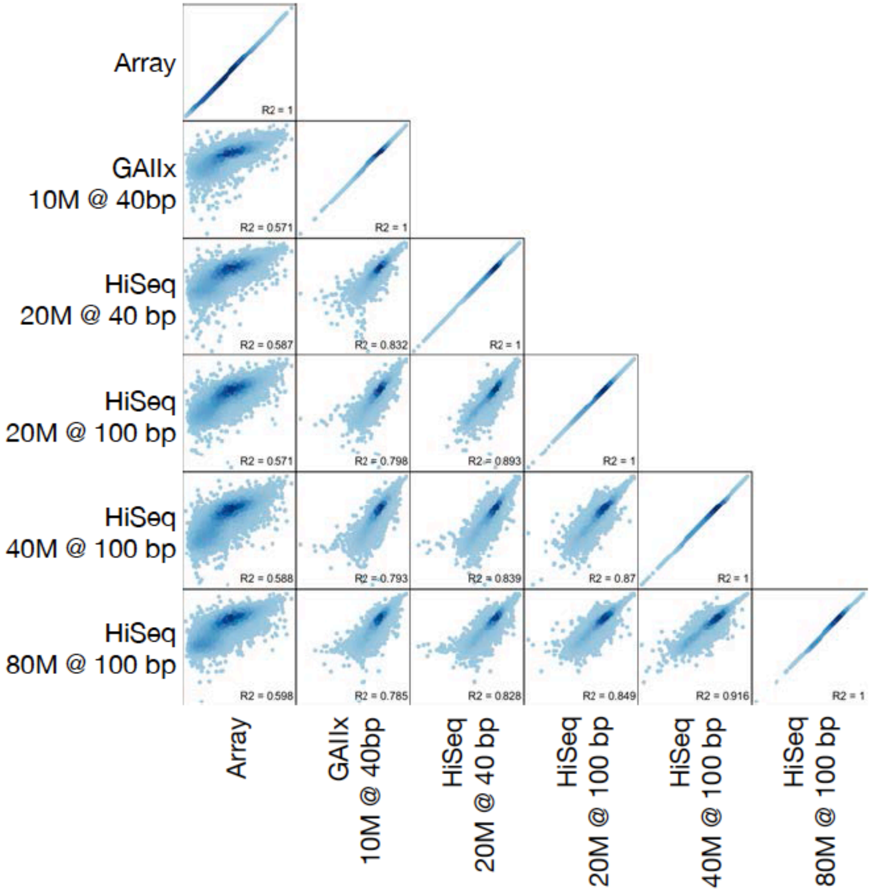
Figure 2 | **Quantifying expression levels: RNA-Seq and microarray compared.** Expression levels are shown, as measured by RNA-Seq and tiling arrays, for *Saccharomyces cerevisiae* cells grown in nutrient-rich media. The two methods agree fairly well for genes with medium levels of expression (middle), but correlation is very low for genes with either low or high expression levels. The tiling array data used in this figure is taken from REF. 2, and the RNA-Seq data is taken from REF. 18.

A. Comparison of Array vs GAIIx



B. Comparison of Array, GAIIx, HiSeq-2000

Platform	Number of Reads	Read Length (in bp)	Number of Genes (coverage > 1)	Number of Transcripts (coverage > 1)
GAIIx	10,583,904	40	8,765	10,460
HiSeq	19,287,803	40	10,541	13,350
HiSeq	19,287,803	100	11,705	16,168
HiSeq	39,392,289	100	12,753	18,462
HiSeq	79,442,311	100	13,465	20,067



Transcriptome Assembly

- *Ab initio* (reference-based)
  - Align reads to genome
    - Unspliced aligners (e.g., BLAT, Bowtie)
    - Splice-aware aligner (e.g., GSNAP, TopHat)
  - Cluster overlapping reads to build a graph
  - Traverse graph to identify isoforms

    (e.g., Cufflinks, Scripture)
- *De novo* (without reference)
  - De Bruijn graph-based approach (e.g., transAbyss)
- Combined strategy
  - e.g., reference genome quality, major rearrangements – cancer cells
- Assembly Quality
  - Use set of well-established transcripts for comparison (different expression levels)

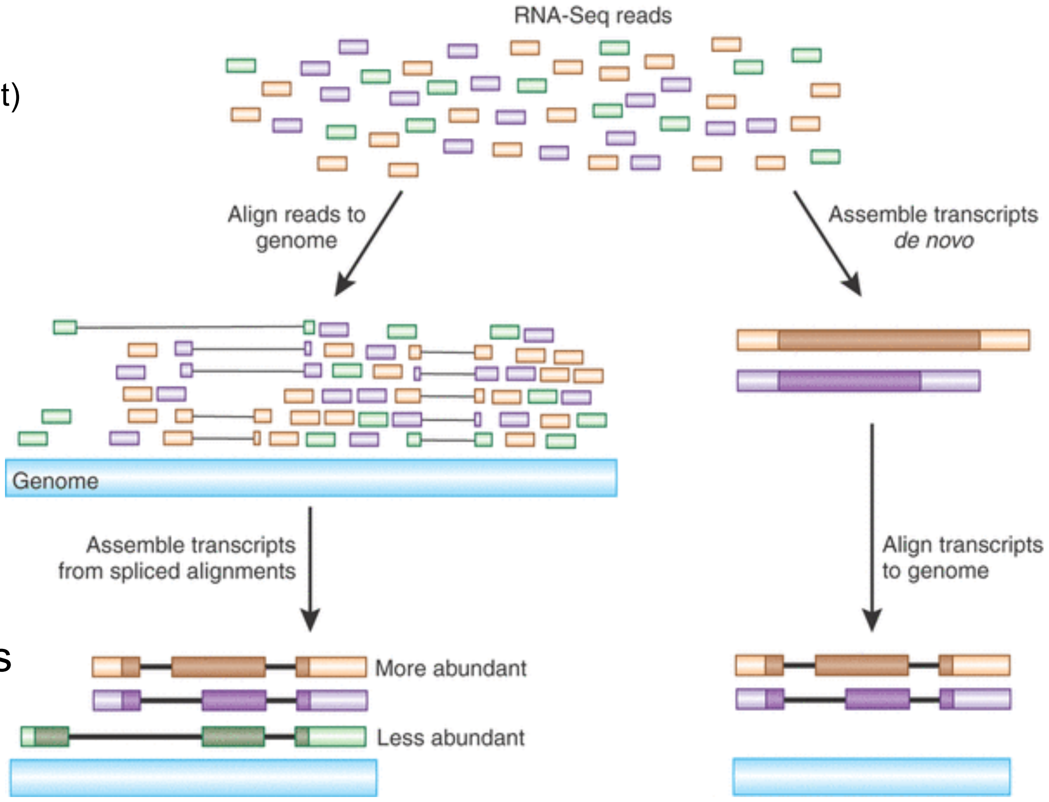
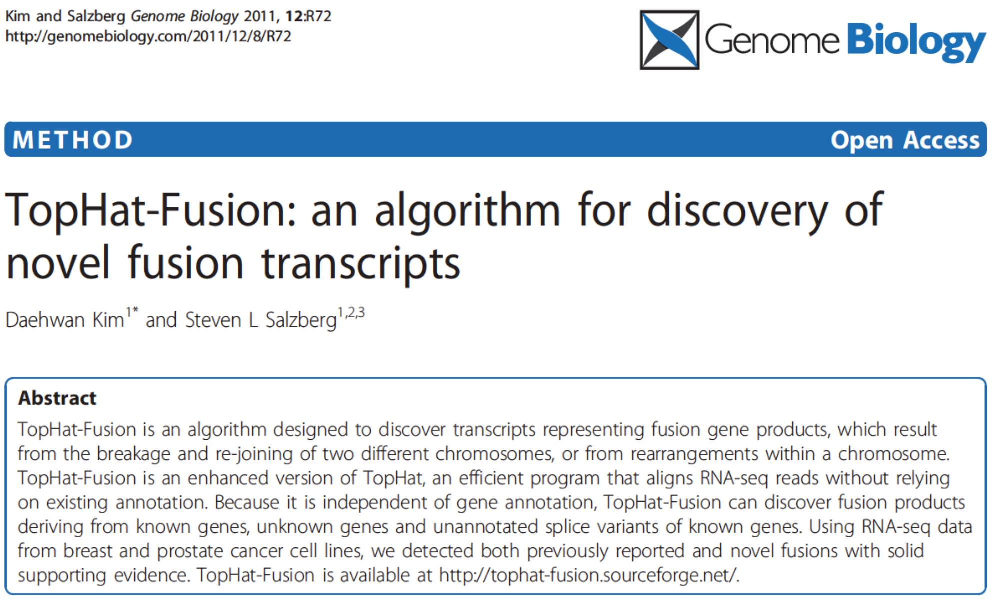

RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Figure 1 Haas & Zody (2010) *Nature Biotechnology*

TopHat-Fusion



Kim and Salzberg *Genome Biology* 2011, **12**:R72
http://genomebiology.com/2011/12/8/R72

Genome **Biology**

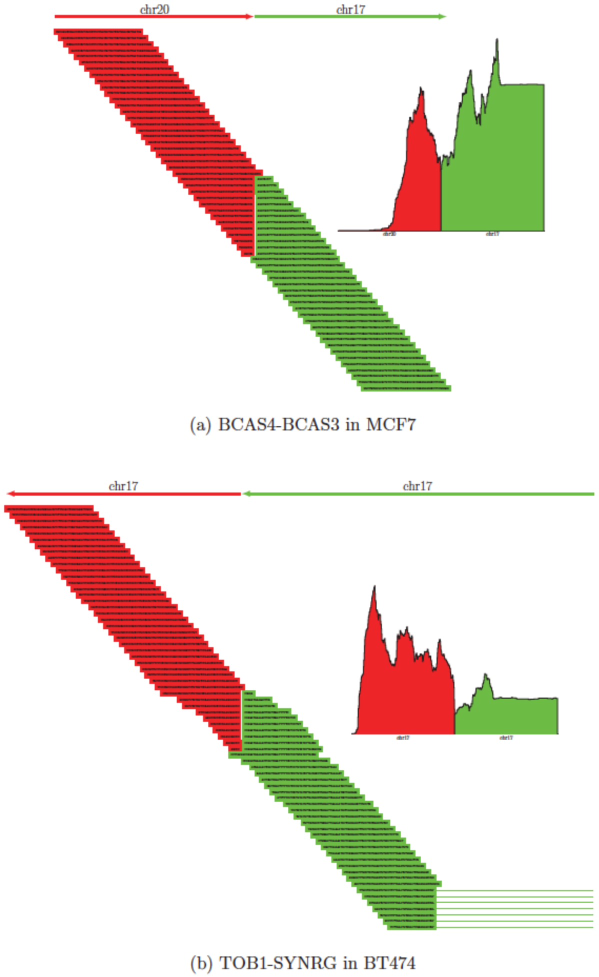
METHOD **Open Access**

TopHat-Fusion: an algorithm for discovery of novel fusion transcripts

Daehwan Kim[1*] and Steven L Salzberg[1,2,3]

Abstract

TopHat-Fusion is an algorithm designed to discover transcripts representing fusion gene products, which result from the breakage and re-joining of two different chromosomes, or from rearrangements within a chromosome. TopHat-Fusion is an enhanced version of TopHat, an efficient program that aligns RNA-seq reads without relying on existing annotation. Because it is independent of gene annotation, TopHat-Fusion can discover fusion products deriving from known genes, unknown genes and unannotated splice variants of known genes. Using RNA-seq data from breast and prostate cancer cell lines, we detected both previously reported and novel fusions with solid supporting evidence. TopHat-Fusion is available at http://tophat-fusion.sourceforge.net/.



(a) BCAS4-BCAS3 in MCF7

(b) TOB1-SYNRG in BT474

Figure 1 Read distributions around two fusions: *BCAS4-BCAS3* and *TOB1-SYNRG*. (a) Sixty reads aligned by TopHat-Fusion that identify a fusion product formed by the *BCAS4* gene on chromosome 20 and the *BCAS3* gene on chromosome 17. The data contained more reads than shown; they are collapsed to illustrate how well they are distributed. The inset figures show the coverage depth in 600-bp windows around each fusion. **(b)** *TOB1* (ENSG0000141232)-*SYNRG* is a novel fusion gene found by TopHat-Fusion, shown here with 70 reads mapping across the fusion point. Note that some of the reads in green span an intron (indicated by thin horizontal lines extending to the right), a feature that can be detected by TopHat's spliced alignment procedure.

(Cited > 450)

ChIP-seq

- ChIP (Chromatin-immunoprecipitation) + next-generation sequencing (NGS)
- **Goals:** Map protein-DNA interactions genome-wide
    - RNA polymerase function
    - transcription factor binding
    - histone modifications
    - nucleosome positioning

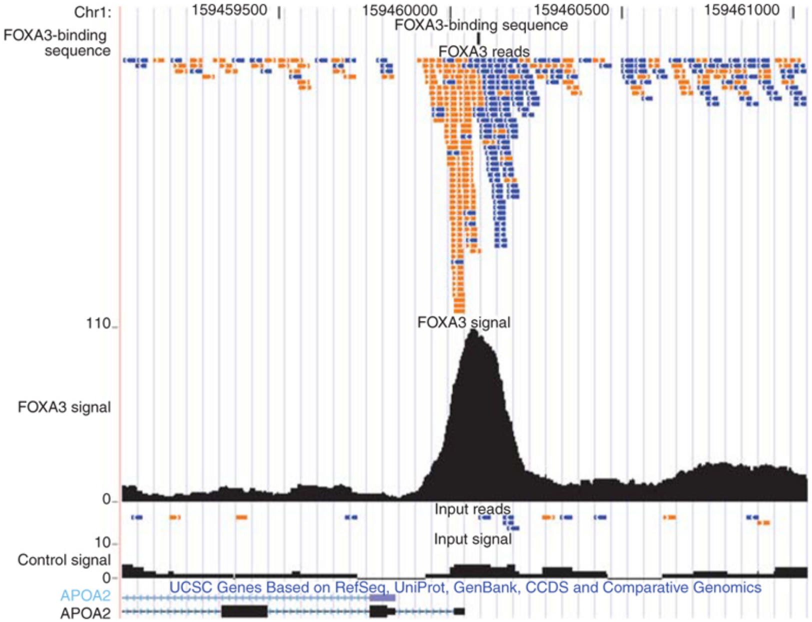
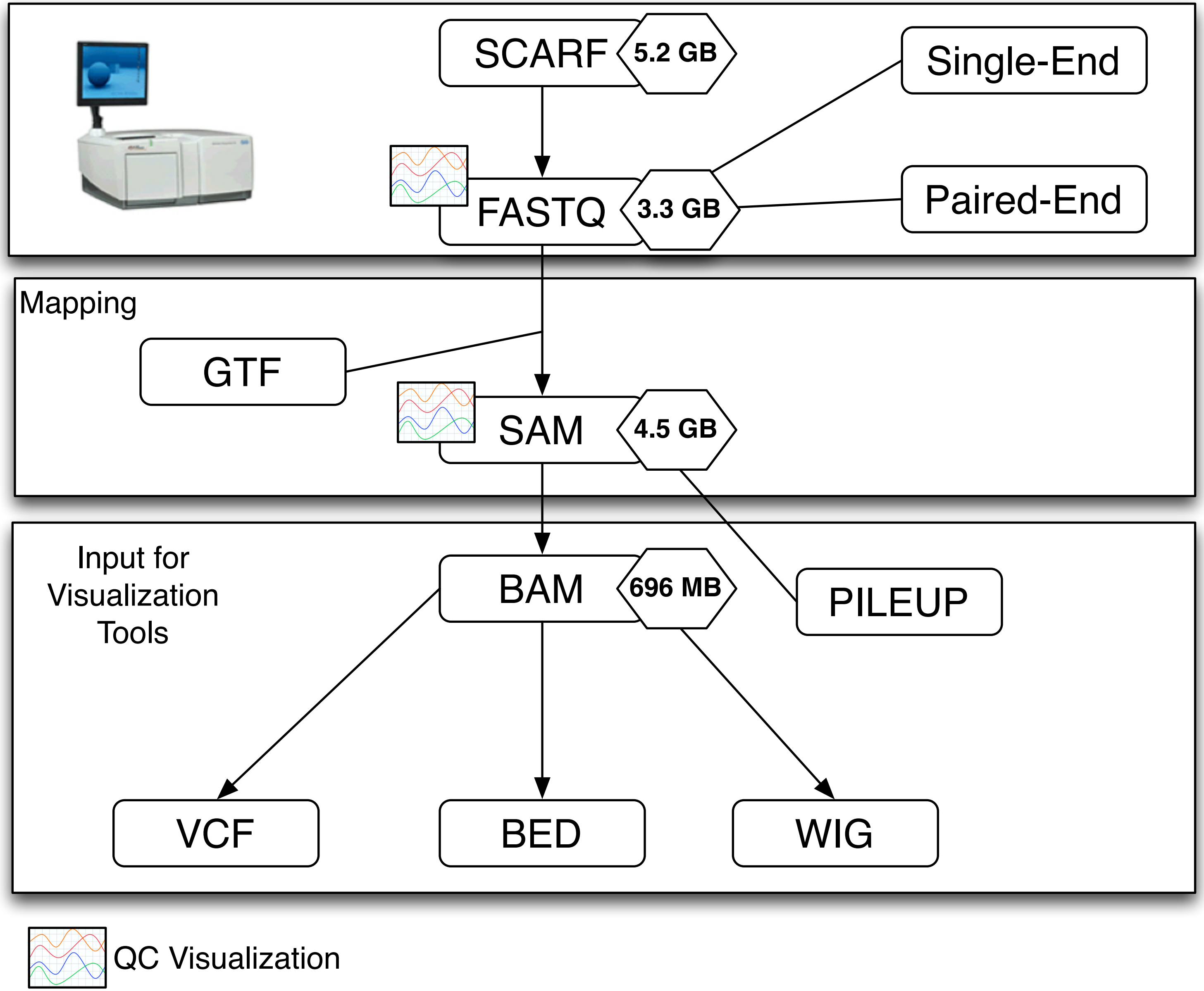


Figure 2 Shah (2009) *Nature Methods*

Park (2009) *Nature Reviews Genetics* 10:669-680
Pepke et al. (2009) *Nature Methods* 6:S22-S32
Leleu et al. (2010) *Brief in Funct Genomics* 9:466-76
Ma & Wong (2011*) Methods in Enzymol* 497:51-73

File Format



SCARF 5.2 GB

Single-End

FASTQ 3.3 GB

Paired-End

Mapping

GTF

SAM 4.5 GB

Input for
Visualization
Tools

BAM 696 MB

PILEUP

VCF

BED

WIG

QC Visualization

Illumina Sequencing Output

(scarf format)

```
@HISEQ:64:C1VDJACXX:5:1:2:836#0/1:CATACAAGTTGTTTGTACTATAGNTGTTTTTGAATT:aabaaaa^abaaba^_]_aaaXPD\^_aaa`Y]_aa
@HISEQ:64:C1VDJACXX:5:1:2:717#0/1:TCTGTTCCAGATTCTAAGGGCATNGTCTTTTTGAAT:aa^]]`\_^[Y_`^aZP^VZV[SDLZ^aa__^^\Ya
@HISEQ:64:C1VDJACXX:5:1:2:188#0/1:TAAGAAGAAAGATGCATAGGTACNATATTTTTGAAT:a``Z[^Y^`\\\^[\^][WNTWNDS_[^_^^[OWY_
@HISEQ:64:C1VDJACXX:5:1:2:1262#0/1:CACTTACAAACAAGGAATGTTGGNCGGTTTTTGAAT:a`ababaabaaaa_``aa``_ULDXZ_^aaa`O_aa
@HISEQ:64:C1VDJACXX:5:1:2:1046#0/1:CTAAGATGGCCTAAGAGTAGACTNACTTTTTTGAAT:abb`Xa`Z_aabaaa`]__Z^`\D\`aaaaaa^aab
@HISEQ:64:C1VDJACXX:5:1:2:748#0/1:CTACATAACATAGAAGTTGGATTNCTCTTTTTGAAT:abba_b`abaaaa\^a``\SQ[OD[aVabaaa_aaa
@HISEQ:64:C1VDJACXX:5:1:2:221#0/1:ATTTCTTGACTTGGATAGAGTTANGTATTTTTGAAT:abba`a`W]^aa]XYa\^TTZ_NDTZ[aaaa_NX]a
@HISEQ:64:C1VDJACXX:5:1:2:664#0/1:CTAACTAGATAGAACTTTGGGGANAAATTTTTGAAT:abbbab\bbba^`a`bV`^``]ZDV]]abaa^X^aa
@HISEQ:64:C1VDJACXXn:5:1:2:197#0/1:CTTCTAGCCCTGGTTTGGGCAGCNGATTTTTGAATT:a_Q^abbaa_b`^``aU_aaaaUDNO^bbab_``Yb
@HISEQ:64:C1VDJACXX:5:1:2:1391#0/1:ATAACTGAGATAAGCTACCGAACNAACTTTTTTAAT:ab`aaa`aaaaa]_`aaa`Y^`RD[___^^`XGQZ_
@HISEQ:64:C1VDJACXX:5:1:2:561#0/1:CACTTCCATCCCAAGTCGTAGCCNAGAGTTTTTGAA:ababab`abaaaaaaa`aaaV_XD[YZX[aaa_O[_
```

(FASTQ format)

```
@HISEQ:64:C1VDJACXX:5:1:2:836#0/1
CATACAAGTTGTTTGTACTATAGNTGTTTTTGAATT
+
aabaaaa^abaaba^_]_aaaXPD\^_aaa`Y]_aa
@HISEQ:64:C1VDJACXX:5:1:2:717#0/1
TCTGTTCCAGATTCTAAGGGCATNGTCTTTTTGAAT
+
aa^]]`\_^[Y_`^aZP^VZV[SDLZ^aa__^^\Ya
@HISEQ:64:C1VDJACXX:5:1:2:188#0/1
TAAGAAGAAAGATGCATAGGTACNATATTTTTGAAT
+
a``Z[^Y^`\\\^[\^][WNTWNDS_[^_^^[OWY_
@HISEQ:64:C1VDJACXX:5:1:2:1262#0/1
CACTTACAAACAAGGAATGTTGGNCGGTTTTTGAAT
+
a`ababaabaaaa_``aa``_ULDXZ_^aaa`O_aa
@HISEQ:64:C1VDJACXX:5:1:2:1046#0/1
CTAAGATGGCCTAAGAGTAGACTNACTTTTTTGAAT
+
abb`Xa`Z_aabaaa`]__Z^`\D\`aaaaaa^aab
```

FASTQ Format

unknown	the unique instrument name
5	flowcell lane
1	tile number within the flowcell lane
2	'x'-coordinate of the cluster within the tile
717	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)p*

Sequence:

```
@HISEQ:64:C1VDJACXX:5:1:2:717#0/1:
TCTGTTCCAGATTCTAAGGGCATNGTCTTTTTGAAT
+
```

Q_{Solexa} :

```
aa^]]`\_^[Y_`^aZP^VZV[SDLZ^aa__^^\Ya
```

```
@unknown_5_1_2_717#0/1
```

Sequence:

```
TCTGTTCCAGATTCTAAGGGCATAGTCTTTTGAATT
+
```

Q_{phred} :

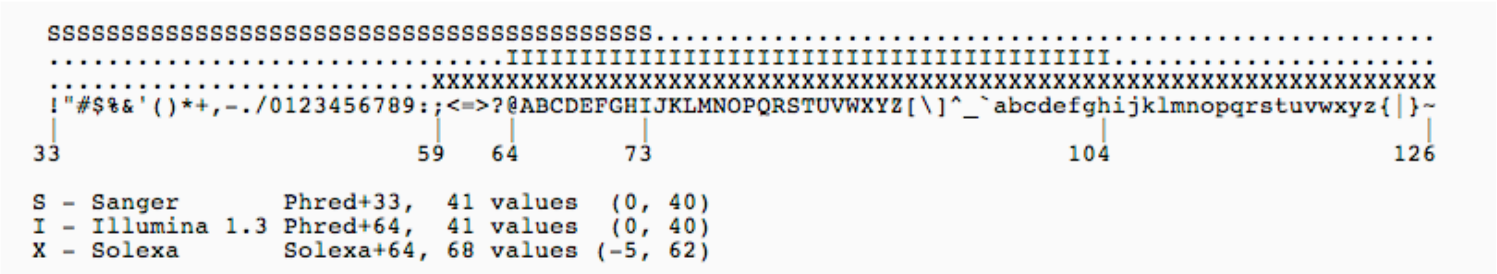
```
33 33 30 29 29 32 28 31 30 27 25 31 32 30 33 26 16 30 22 26 22 27 19 5 12 26 30
```

(Worst) $0 \le Q_{phred} \le 40$ (Perfect)

Quality Score

A quality value Q is an integer mapping of *p* (i.e., the probability that the corresponding base call is incorrect).

ASCII

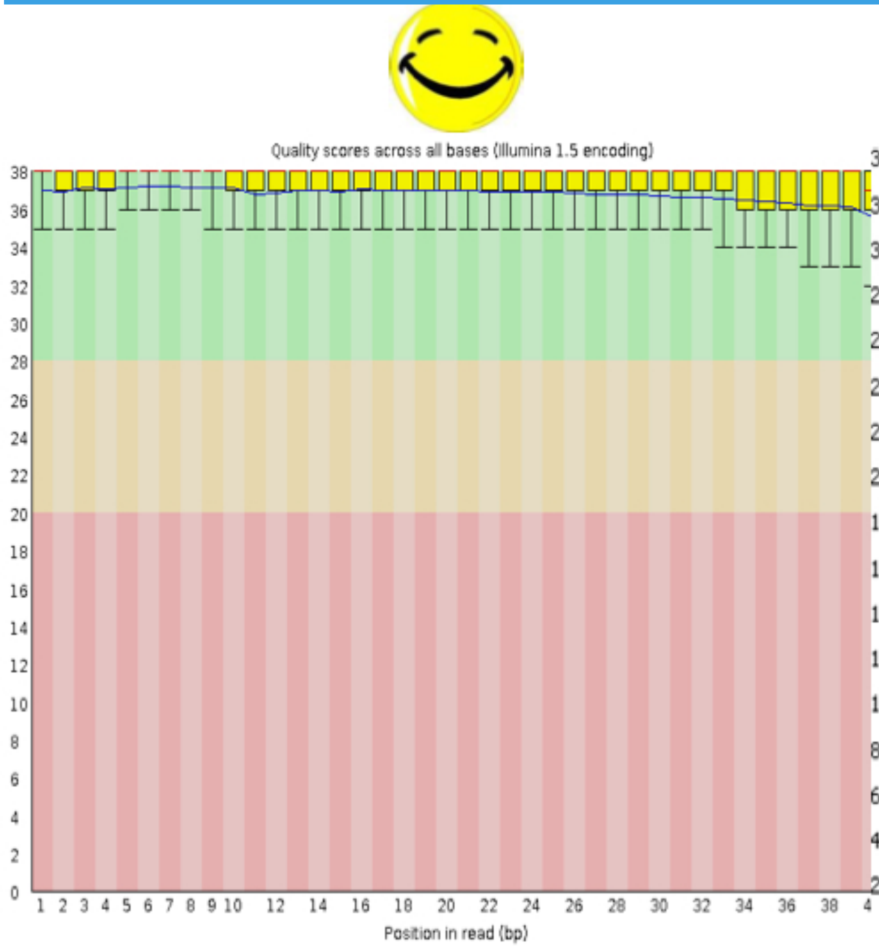


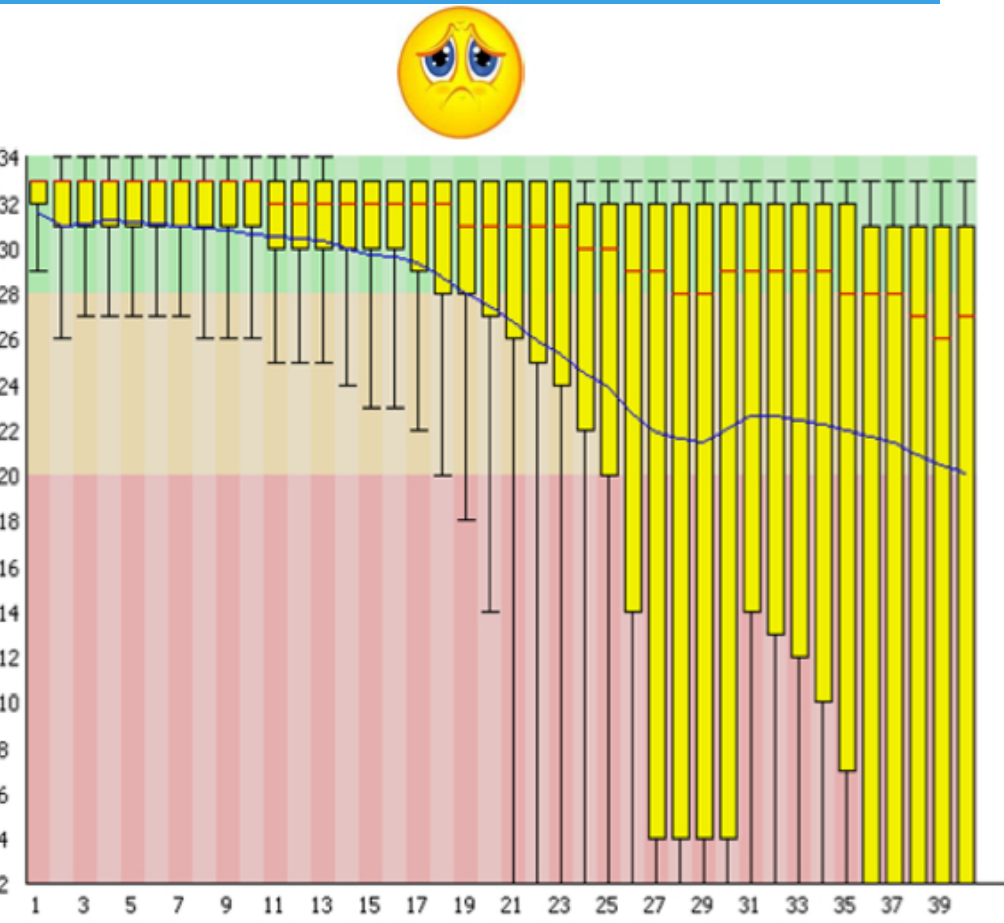
```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger       Phred+33,  41 values  (0, 40)
I - Illumina 1.3 Phred+64,  41 values  (0, 40)
X - Solexa       Solexa+64, 68 values (-5, 62)
```

- Illumina 1.3+ FASTQ :
  - Phred scores with ASCII offset of 64
  - Phred scores from 0 to 62
  - From the raw score *p*(scarf file),
    Q_{Solexa} = ascii(*p*) – 64

$$Q_{\text{PHRED}} = 10 \times \log_{10}\left(10^{Q_{\text{Solexa}}/10} + 1\right)$$

FASTQ QC Visualization
Per base sequence quality

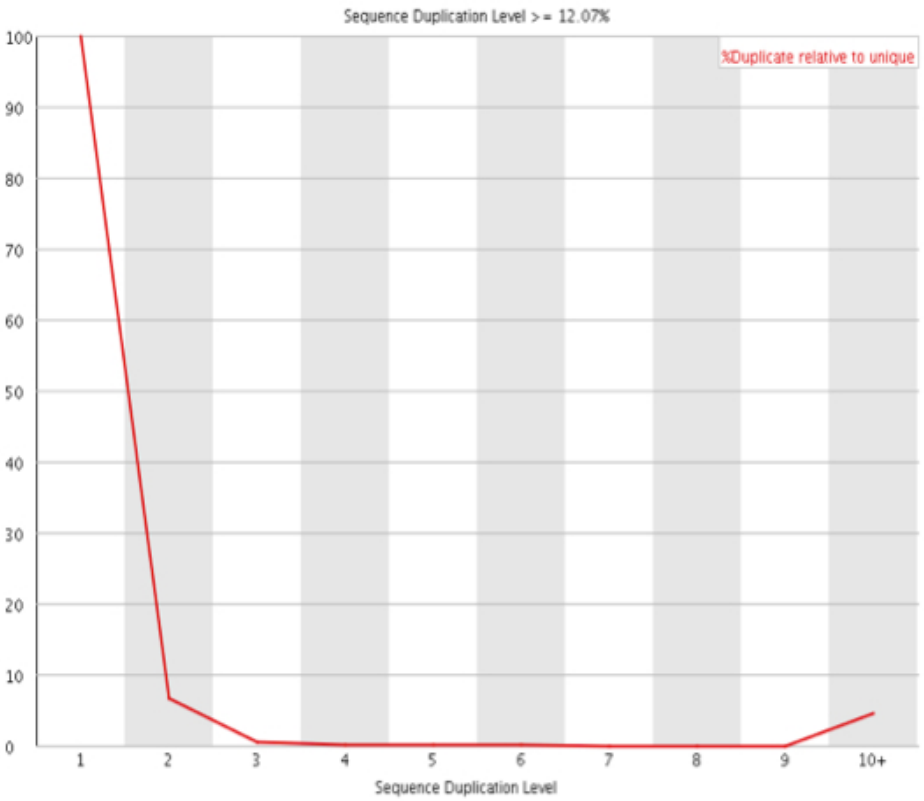


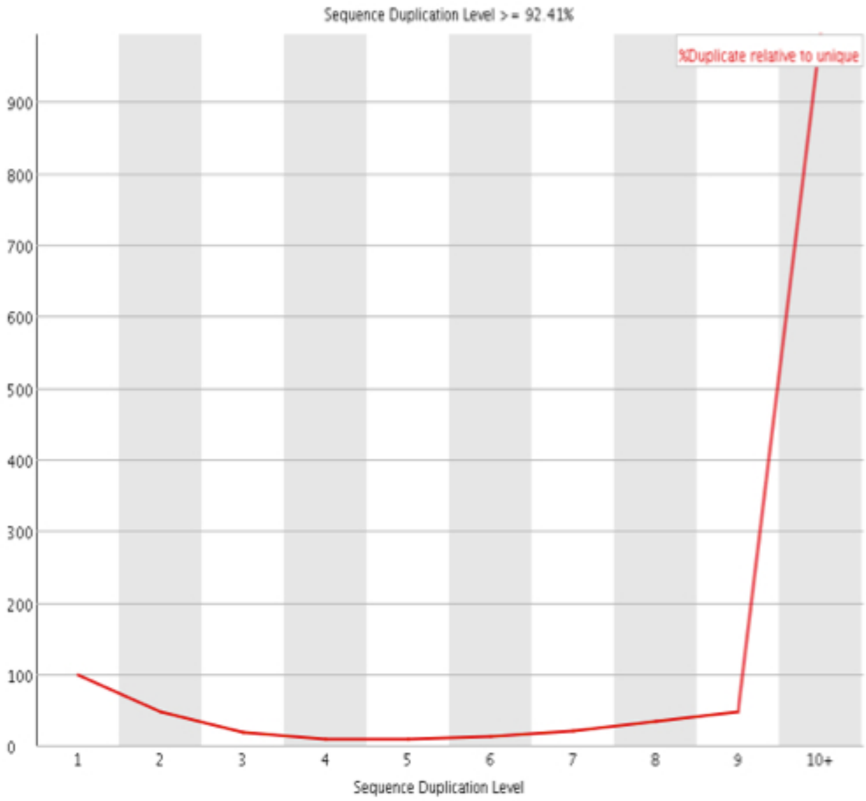


http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

(Slides from Tzu Phang)

Duplication Level

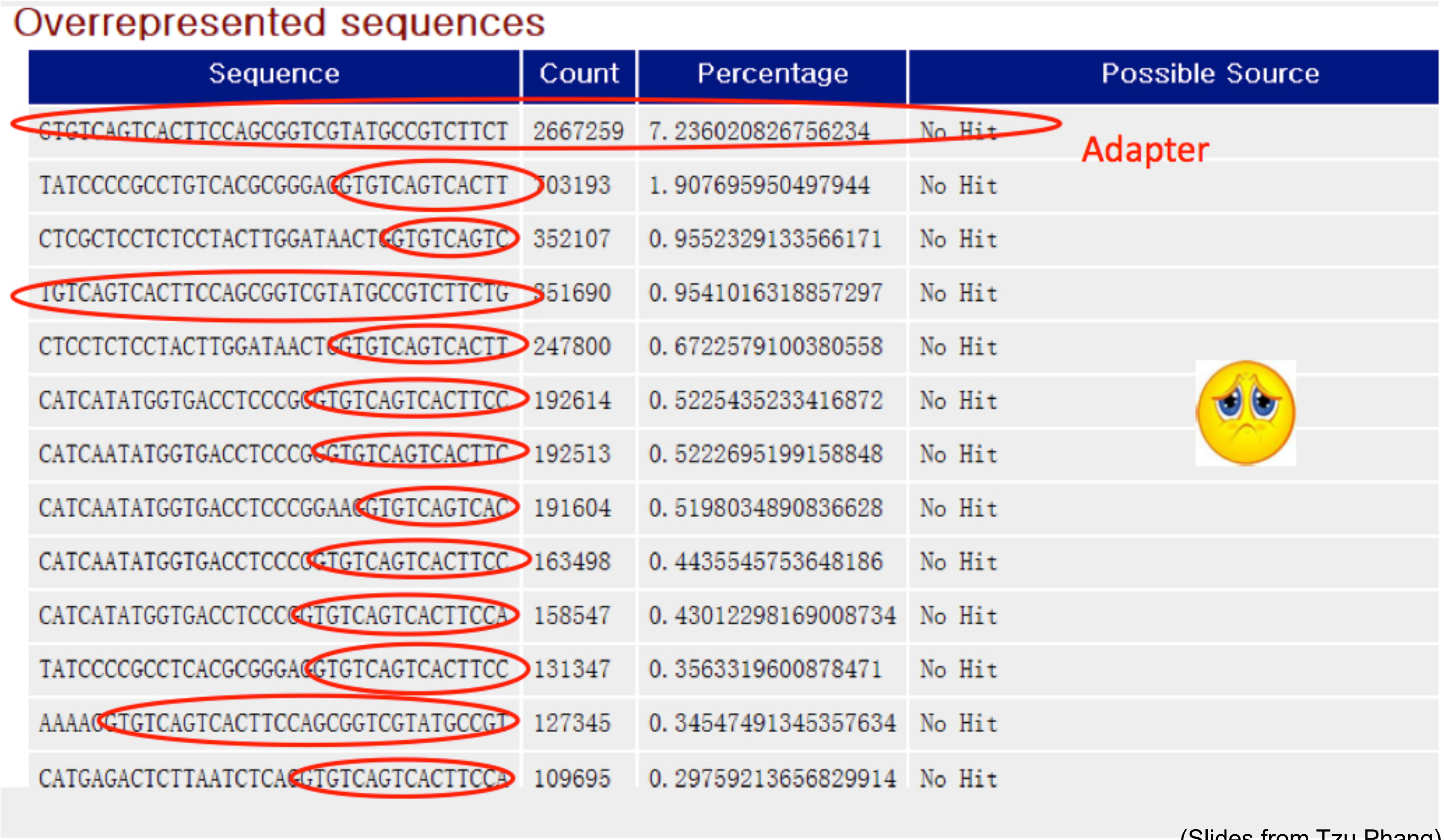




Sequence Characteristic / Duplication Level

(Slides from Tzu Phang)

Over-represented Sequences



Overrepresented sequences

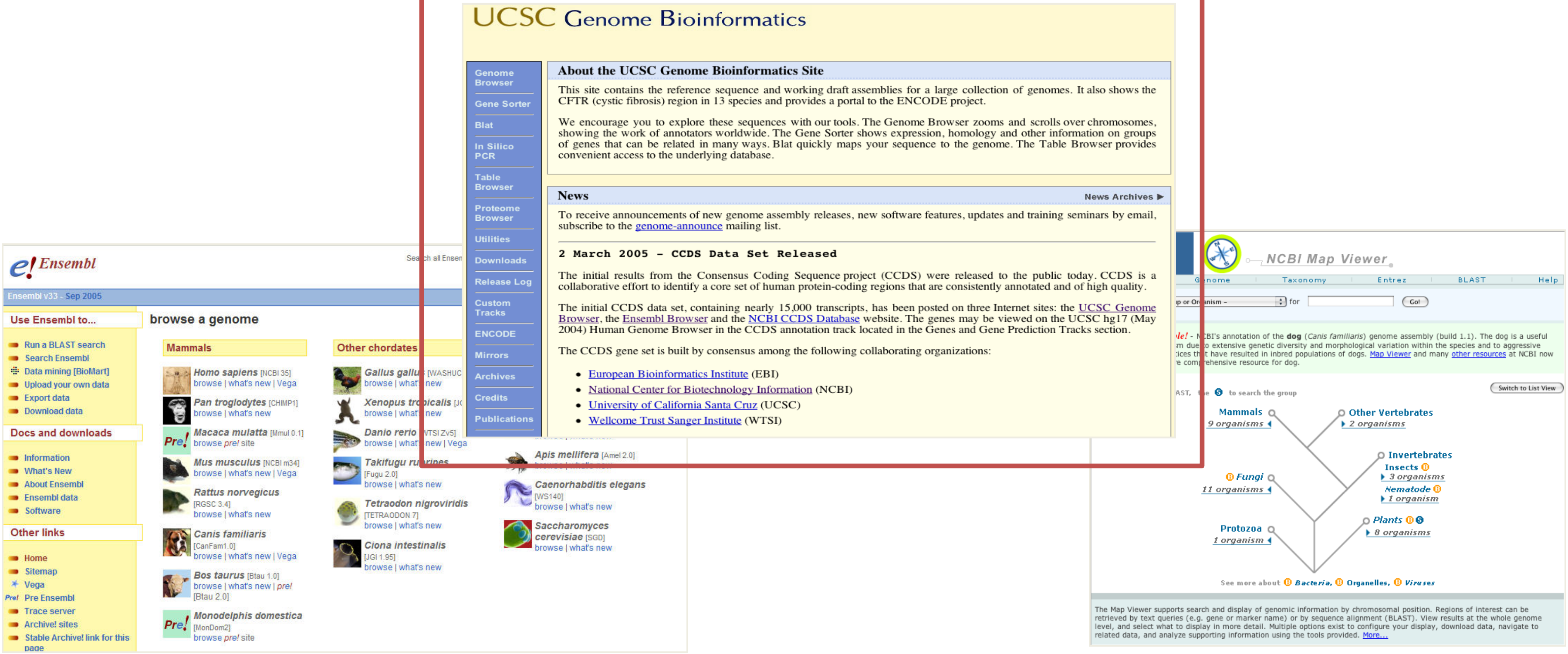
Sequence	Count	Percentage	Possible Source
GTGTCAGTCACTTCCAGCGGTCGTATGCCGTCTTCT	2667259	7.236020826756234	No Hit
TATCCCCGCCTGTCACGCGGGAGGTGTCAGTCACTT	703193	1.907695950497944	No Hit
CTCGCTCCTCTCCTACTTGGATAACTGGTGTCAGTC	352107	0.9552329133566171	No Hit
TGTCAGTCACTTCCAGCGGTCGTATGCCGTCTTCTG	351690	0.9541016318857297	No Hit
CTCCTCTCCTACTTGGATAACTGGTGTCAGTCACTT	247800	0.6722579100380558	No Hit
CATCATATGGTGACCTCCCGGGTGTCAGTCACTTCC	192614	0.5225435233416872	No Hit
CATCAATATGGTGACCTCCCGGGTGTCAGTCACTTC	192513	0.5222695199158848	No Hit
CATCAATATGGTGACCTCCCGGAAGGTGTCAGTCAC	191604	0.5198034890836628	No Hit
CATCAATATGGTGACCTCCCGGTGTCAGTCACTTCC	163498	0.4435545753648186	No Hit
CATCATATGGTGACCTCCCGGTGTCAGTCACTTCCA	158547	0.43012298169008734	No Hit
TATCCCCGCCTCACGCGGGAGGTGTCAGTCACTTCC	131347	0.3563319600878471	No Hit
AAAAGGTGTCAGTCACTTCCAGCGGTCGTATGCCGT	127345	0.34547491345357634	No Hit
CATGAGACTCTTAATCTCAGGTGTCAGTCACTTCCA	109695	0.29759213656829914	No Hit

Adapter

(Slides from Tzu Phang)

Genome Browsers

UCSC Genome Browser



About the UCSC Genome Bioinformatics Site

This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also shows the CFTR (cystic fibrosis) region in 13 species and provides a portal to the ENCODE project.

We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database.

News

To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list.

2 March 2005 - CCDS Data Set Released

The initial results from the Consensus Coding Sequence project (CCDS) were released to the public today. CCDS is a collaborative effort to identify a core set of human protein-coding regions that are consistently annotated and of high quality.

The initial CCDS data set, containing nearly 15,000 transcripts, has been posted on three Internet sites: the UCSC Genome Browser, the Ensembl Browser and the NCBI CCDS Database website. The genes may be viewed on the UCSC hg17 (May 2004) Human Genome Browser in the CCDS annotation track located in the Genes and Gene Prediction Tracks section.

The CCDS gene set is built by consensus among the following collaborating organizations:

- European Bioinformatics Institute (EBI)
- National Center for Biotechnology Information (NCBI)
- University of California Santa Cruz (UCSC)
- Wellcome Trust Sanger Institute (WTSI)

le! - N BI's annotation of the **dog** (*Canis familiaris*) genome assembly (build 1.1). The dog is a useful m due o extensive genetic diversity and morphological variation within the species and to aggressive ices th t have resulted in inbred populations of dogs. Map Viewer and many other resources at NCBI now e com rehensive resource for dog.

The Map Viewer supports search and display of genomic information by chromosomal position. Regions of interest can be retrieved by text queries (e.g. gene or marker name) or by sequence alignment (BLAST). View results at the whole genome level, and select what to display in more detail. Multiple options exist to configure your display, download data, navigate to related data, and analyze supporting information using the tools provided. More...

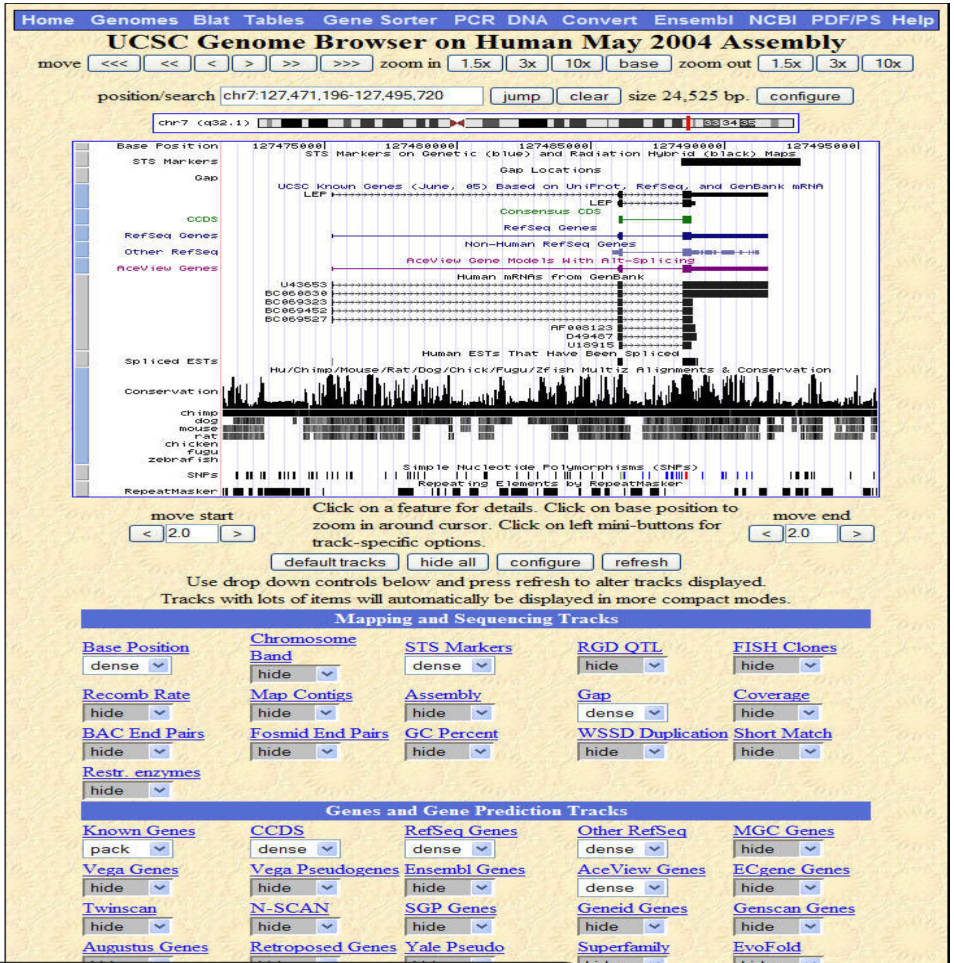
EBI Ensembl

Wang, J. (2013). A brief introduction to web-based genome browsers. *Briefings in bioinformatics*, *14*(2), 131–143

NCBI Map Viewer

(Slides from Tzu Phang)

UCSC Genome Browser



Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Help

UCSC Genome Browser on Human May 2004 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr7:127,471,196-127,495,720 jump clear size 24,525 bp. configure

move start < 2.0 >

Click on a feature for details. Click on base position to zoom in around cursor. Click on left mini-buttons for track-specific options.

move end < 2.0 >

default tracks hide all configure refresh

Use drop down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

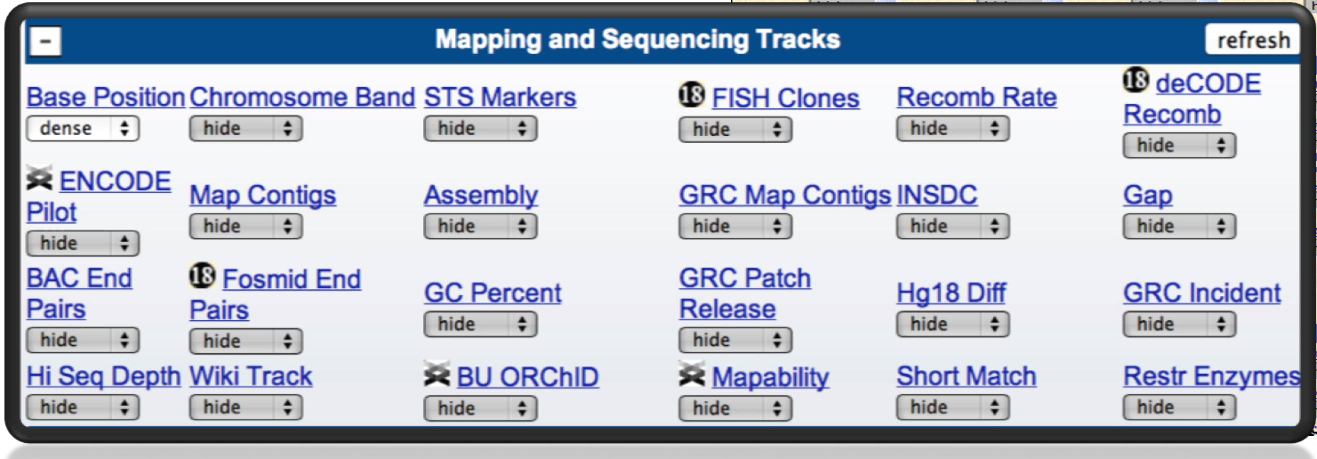
Mapping and Sequencing Tracks

Base Position dense	Chromosome Band hide	STS Markers dense	RGD QTL hide	FISH Clones hide
Recomb Rate hide	Map Contigs hide	Assembly hide	Gap dense	Coverage hide
BAC End Pairs hide	Fosmid End Pairs hide	GC Percent hide	WSSD Duplication hide	Short Match hide
Restr. enzymes hide				

Genes and Gene Prediction Tracks

Known Genes pack	CCDS dense	RefSeq Genes dense	Other RefSeq dense	MGC Genes hide
Vega Genes hide	Vega Pseudogenes hide	Ensembl Genes hide	AceView Genes dense	ECgene Genes hide
Twinscan hide	N-SCAN hide	SGP Genes hide	Geneid Genes hide	Genscan Genes hide
Augustus Genes	Retroposed Genes	Yale Pseudo	Superfamily hide	EvoFold hide

ther mRNAs hide	Other ESTs hide
ene Bounds hide	Alt-Splicing hide
ffy GNF1H hide	Affy U133Plus2 hide
x Reg Potential hide	TFBS Conserved hide
ugu Chain hide	Fugu Net hide
ebrafish Chain hide	Zebrafish Net hide
Chicken Net	Opossum

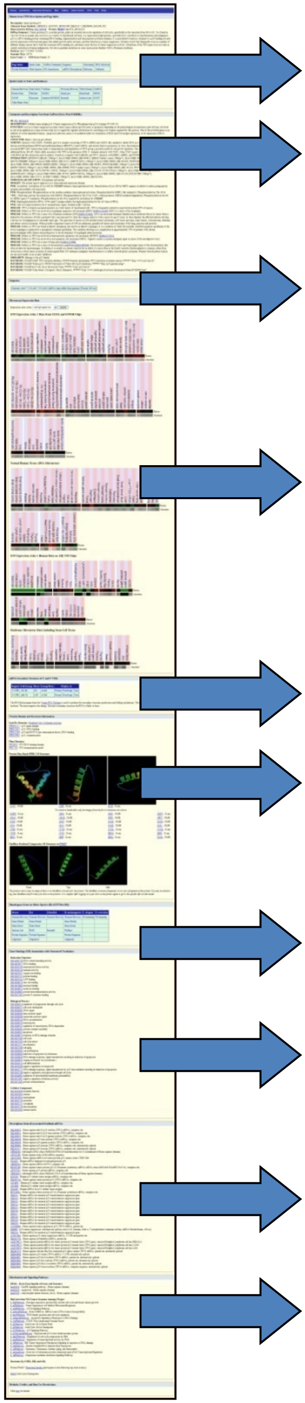


– Mapping and Sequencing Tracks refresh

Base Position dense	Chromosome Band hide	STS Markers hide	FISH Clones hide	Recomb Rate hide	deCODE Recomb hide
ENCODE Pilot hide	Map Contigs hide	Assembly hide	GRC Map Contigs hide	INSDC hide	Gap hide
BAC End Pairs hide	Fosmid End Pairs hide	GC Percent hide	GRC Patch Release hide	Hg18 Diff hide	GRC Incident hide
Hi Seq Depth hide	Wiki Track hide	BU ORChID hide	Mapability hide	Short Match hide	Restr Enzymes hide

(Slides from Tzu Phang)

Click annotation track item for details pages



informative description

other resource links

links to sequences

microarray data

mRNA secondary structure

protein domains/structure

homologs in other species

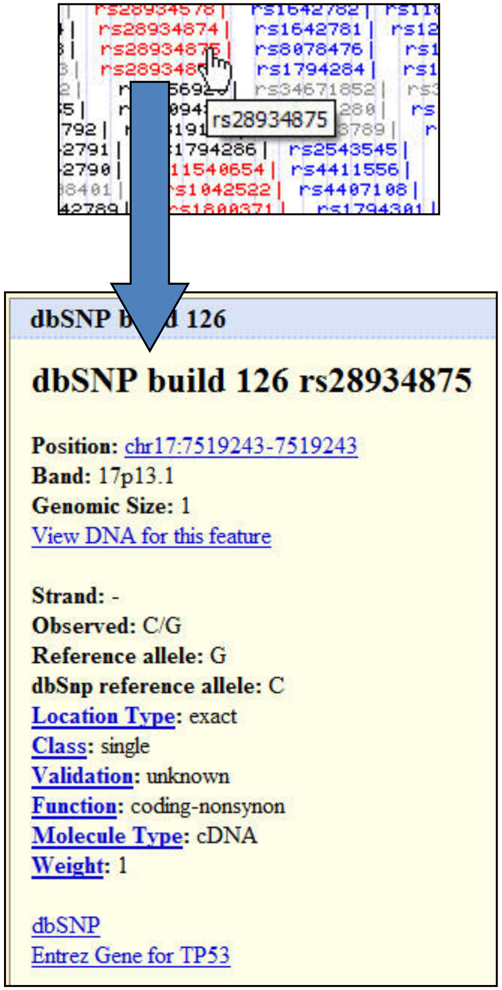
Gene Ontology™ descriptions

mRNA descriptions

pathways

Not all genes have This much detail.

Different annotation tracks carry different data.



dbSNP build 126

dbSNP build 126 rs28934875

Position: chr17:7519243-7519243
Band: 17p13.1
Genomic Size: 1
View DNA for this feature

Strand: -
Observed: C/G
Reference allele: G
dbSnp reference allele: C
Location Type: exact
Class: single
Validation: unknown
Function: coding-nonsynon
Molecule Type: cDNA
Weight: 1

dbSNP
Entrez Gene for TP53

(Slides from Tzu Phang)

UCSC Genome Browser

✔ Pretty of annotation to browse
✔ Not species specific
✔ Retrieve annotation / data

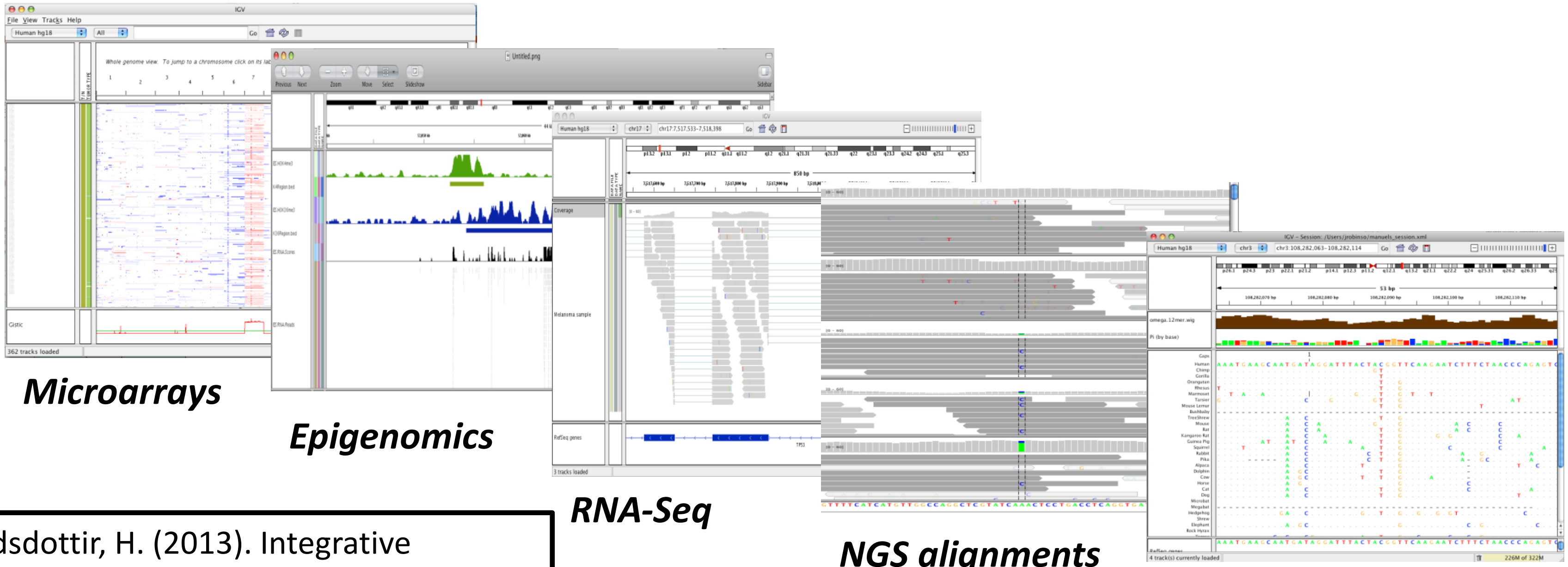
✘ Not dynamic – need refresh
✘ Not NGS data friendly
✘ Graphic render in server, slow

However – still very useful to build customized genome browsers …

(Slides from Tzu Phang)

What is IGV

A desktop application for integrated visualization of multiple data types and annotations in the context of the genome



Microarrays

Epigenomics

RNA-Seq

NGS alignments

Comparative genomics

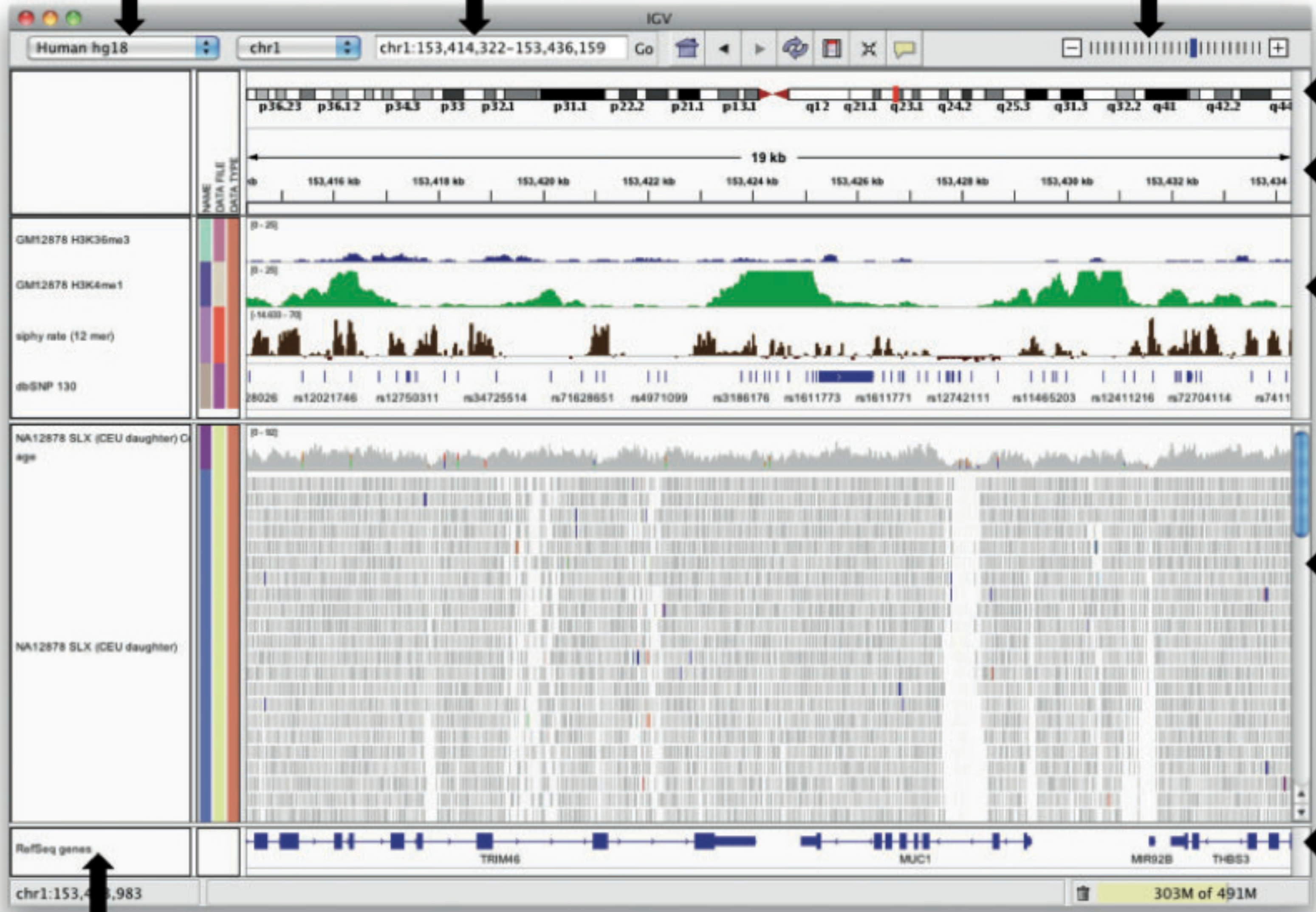
Thorvaldsdottir, H. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, *14*(2), 178–192.

(Slides from Tzu Phang)

Reference Genome Selector

Search Box

Zoom Bar



IGV

Human hg18

chr1

chr1:153,414,322-153,436,159

Go

NAME
DATA FILE
DATA TYPE

GM12878 H3K36me3

GM12878 H3K4me1

siphy rate (12 mer)

dbSNP 130

NA12878 SLX (CEU daughter) Coverage

NA12878 SLX (CEU daughter)

RefSeq genes

chr1:153,4 ,983

303M of 491M

Chromosome Ideogram

Genomic Coordinates

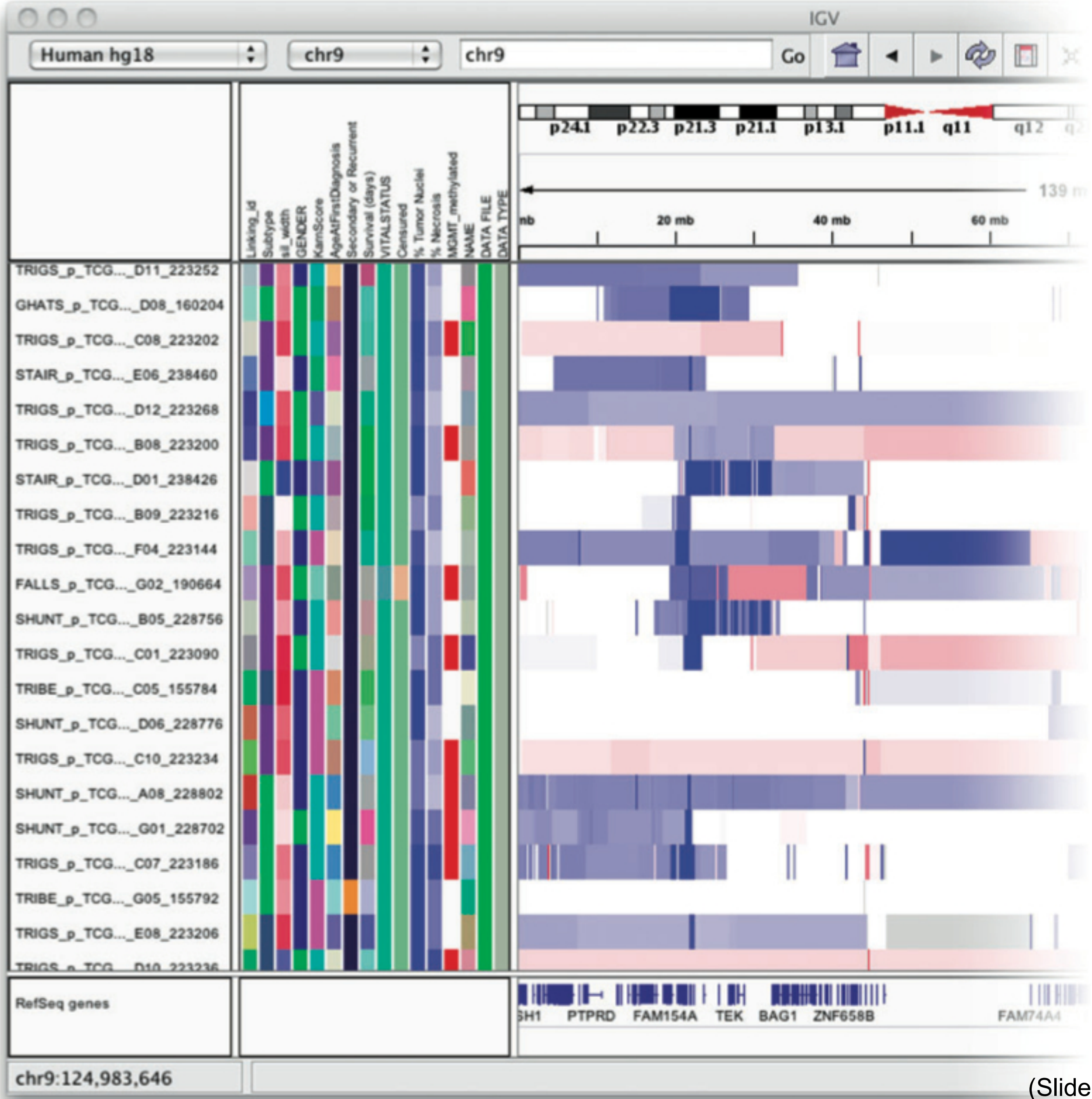
Data tracks (ChIP-seq)

Data tracks (NGS reads)

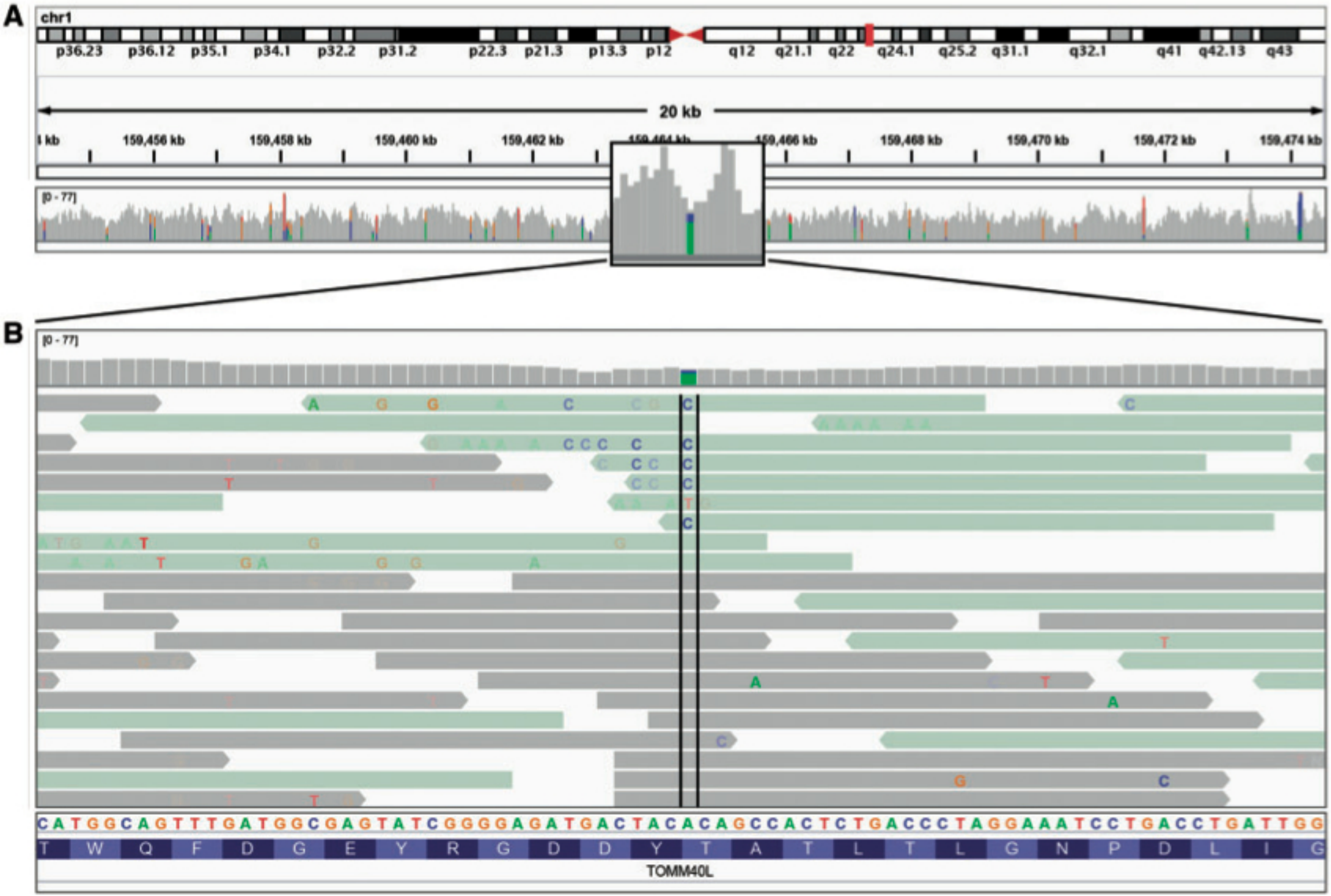
Reference Gene Track

Track Names

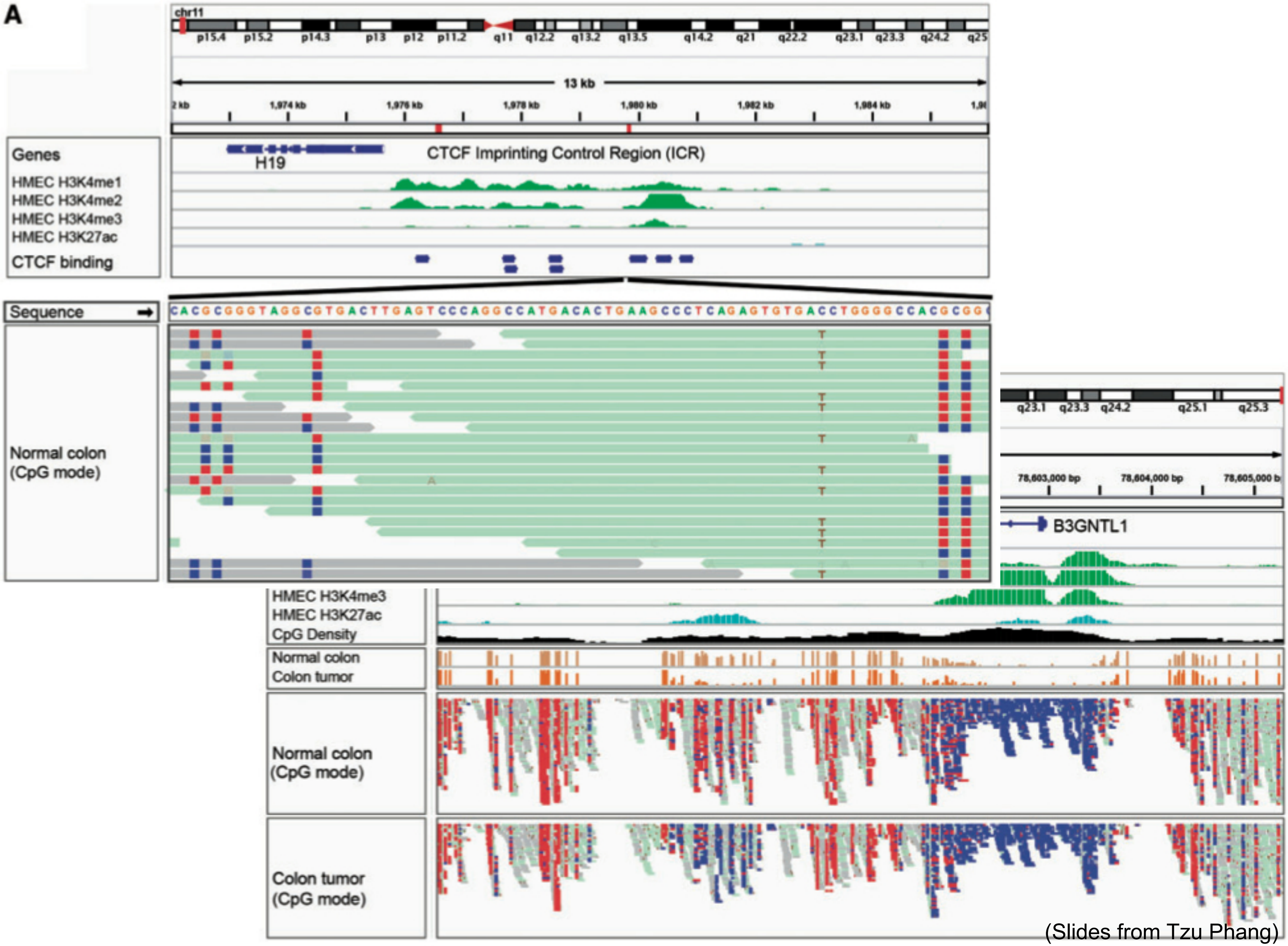
(Slides from Tzu Phang)


IGV
Human hg18
chr9
chr9
Go
p24.1
p22.3
p21.3
p21.1
p13.1
p11.1
q11
q12
139 m
mb
20 mb
40 mb
60 mb
Linking_id
Subtype
sil_width
GENDER
KarnScore
AgeAtFirstDiagnosis
Secondary or Recurrent
Survival (days)
VITALSTATUS
Censured
% Tumor Nuclei
% Necrosis
MGMT_methylated
NAME
DATA FILE
DATA TYPE
TRIGS_p_TCG..._D11_223252
GHATS_p_TCG..._D08_160204
TRIGS_p_TCG..._C08_223202
STAIR_p_TCG..._E06_238460
TRIGS_p_TCG..._D12_223268
TRIGS_p_TCG..._B08_223200
STAIR_p_TCG..._D01_238426
TRIGS_p_TCG..._B09_223216
TRIGS_p_TCG..._F04_223144
FALLS_p_TCG..._G02_190664
SHUNT_p_TCG..._B05_228756
TRIGS_p_TCG..._C01_223090
TRIBE_p_TCG..._C05_155784
SHUNT_p_TCG..._D06_228776
TRIGS_p_TCG..._C10_223234
SHUNT_p_TCG..._A08_228802
SHUNT_p_TCG..._G01_228702
TRIGS_p_TCG..._C07_223186
TRIBE_p_TCG..._G05_155792
TRIGS_p_TCG..._E08_223206
TRIGS_p_TCG..._D10_223236
RefSeq genes
SH1
PTPRD
FAM154A
TEK
BAG1
ZNF658B
FAM74A4
chr9:124,983,646

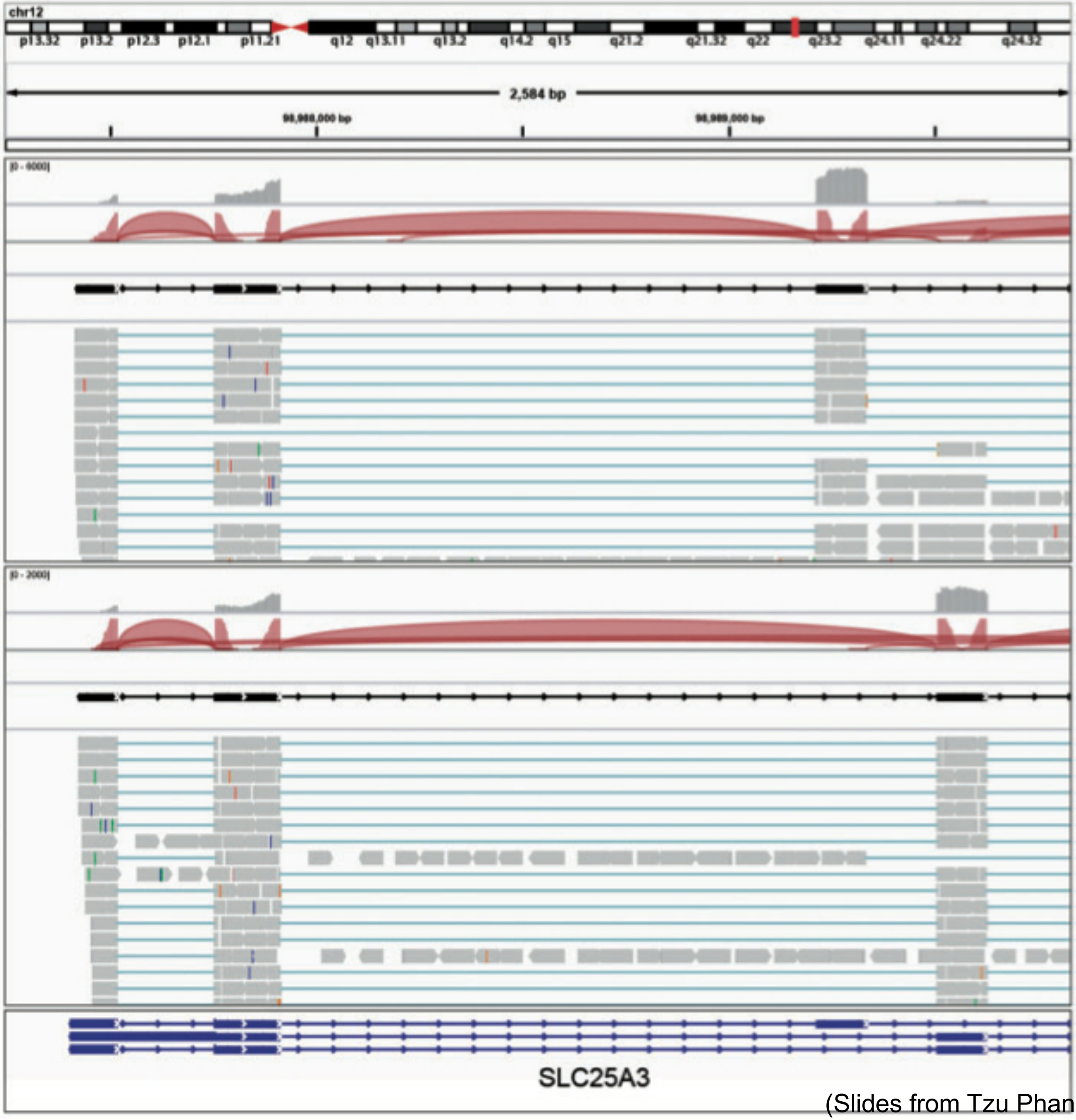
(Slides from Tzu Phang)



(Slides from Tzu Phang)



(Slides from Tzu Phang)



Heart

Coverage

Junctions

Transcripts

Alignments

Liver

Coverage

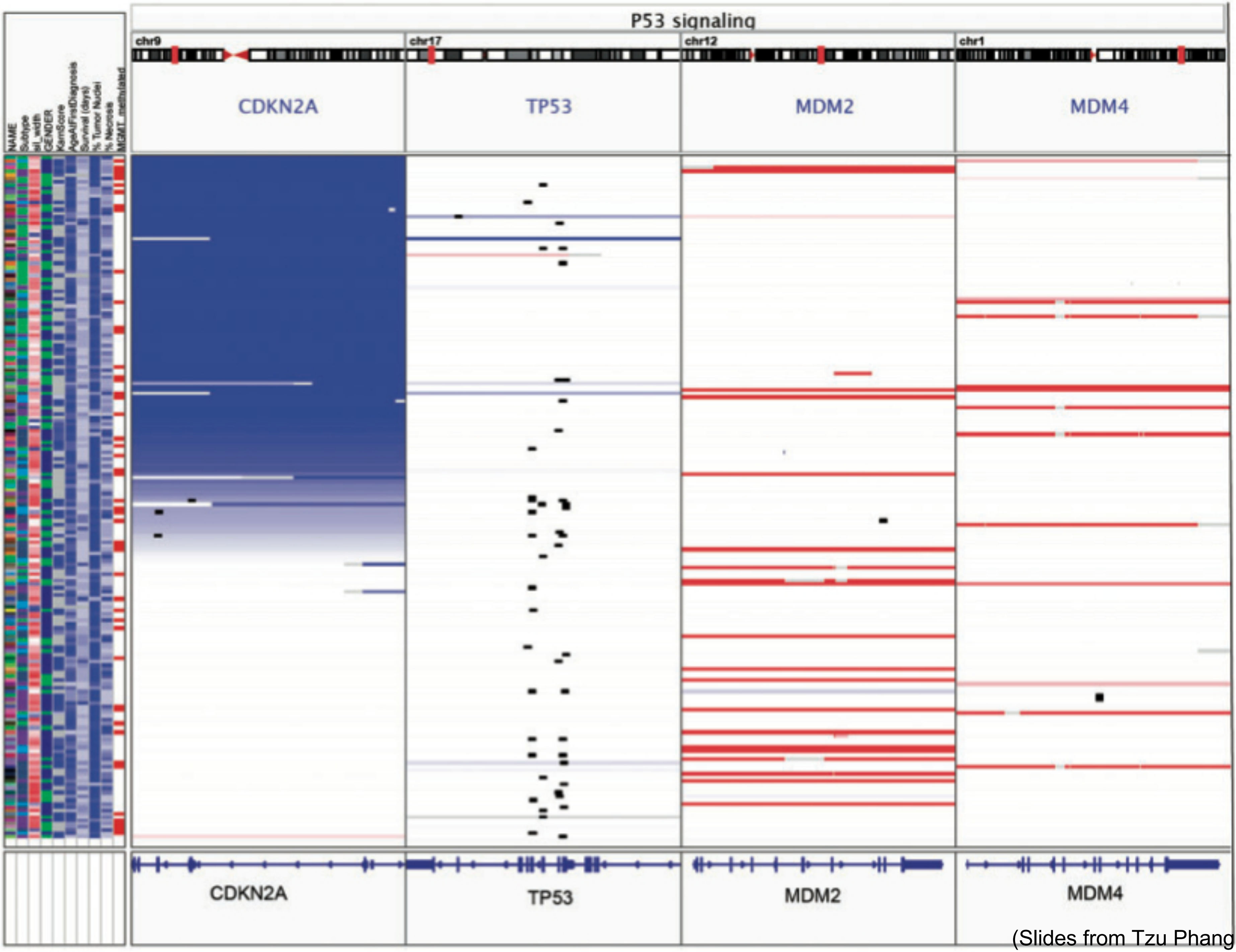
Junctions

Transcripts

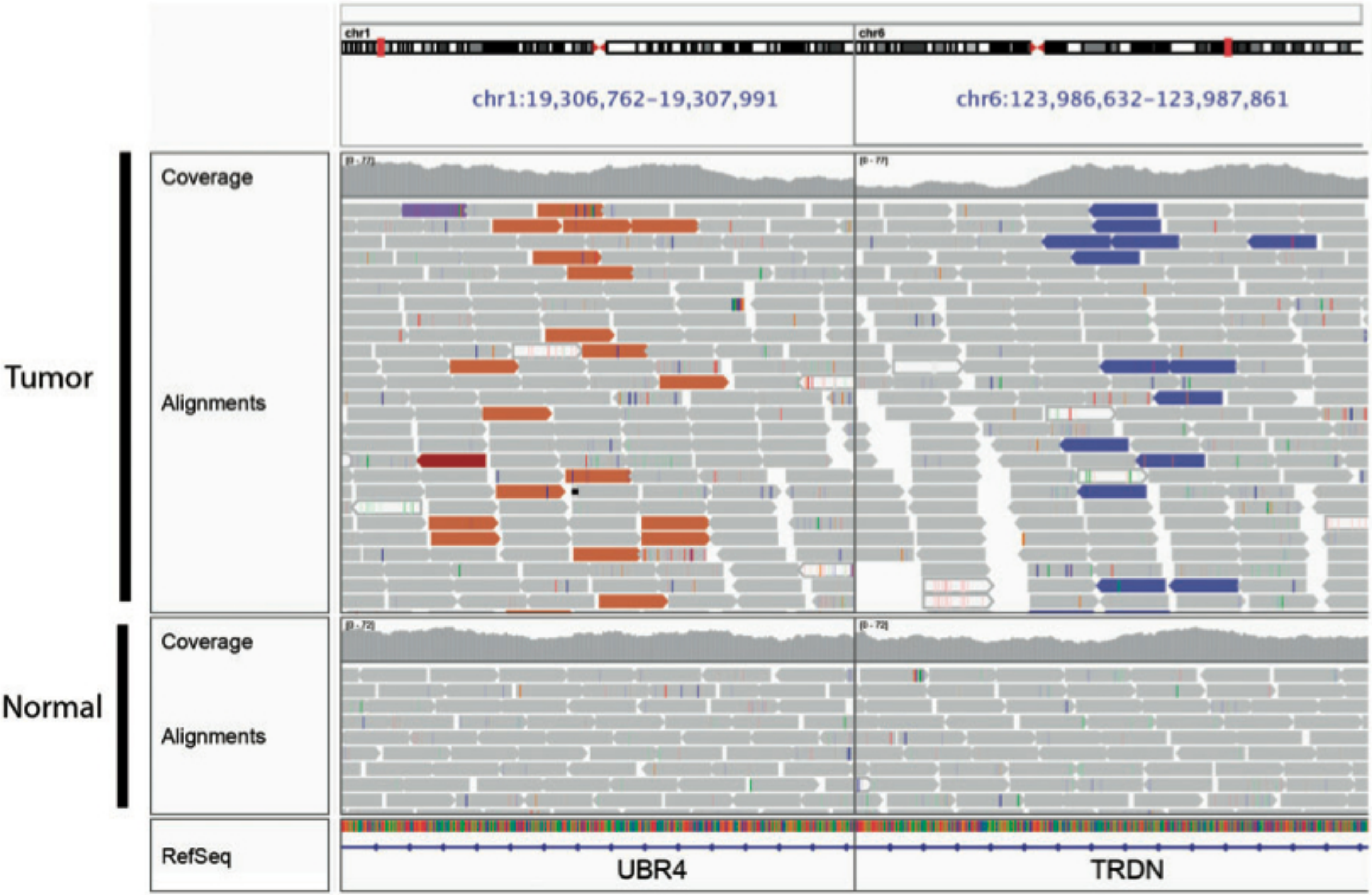
Alignments

Refseq

(Slides from Tzu Phang)



(Slides from Tzu Phang)



(Slides from Tzu Phang)

IGV

✔Locally installed; locally render

✔Very Dynamic; zooming, panning, …

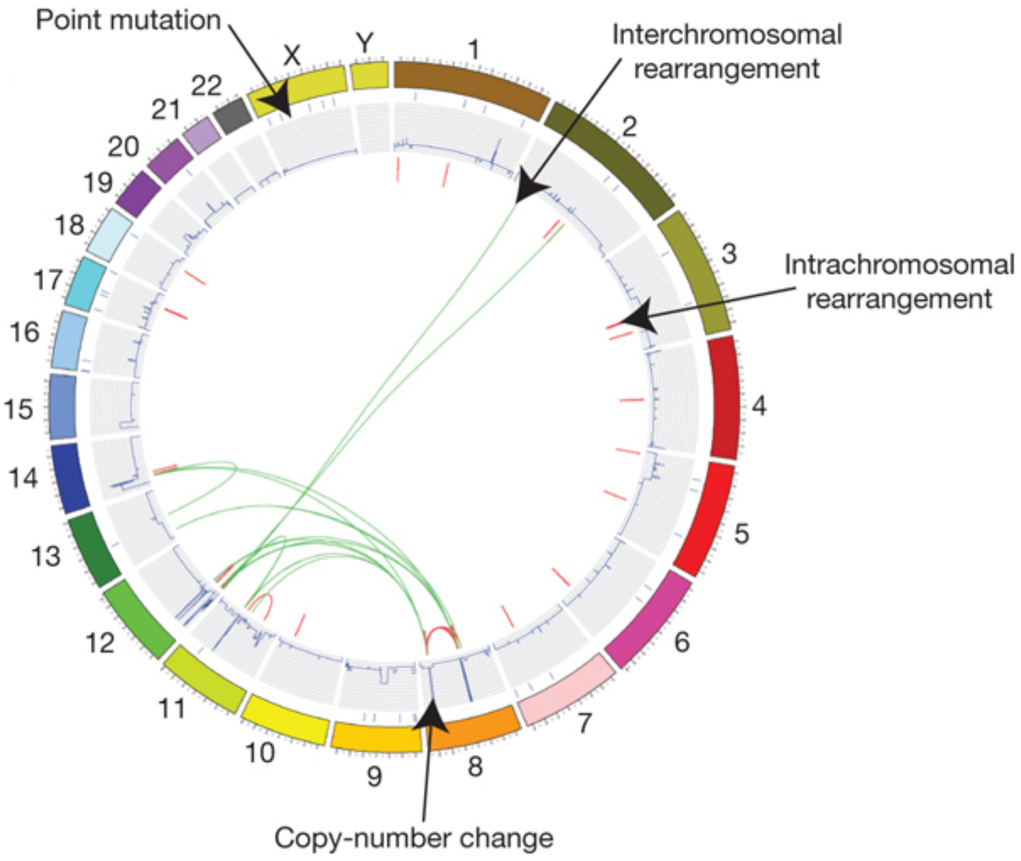
✔Customizable

✔Very NGS data friendly

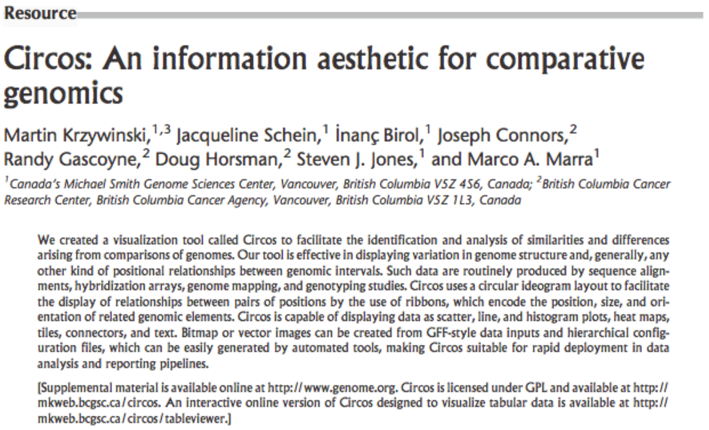
✔Not much analysis

(Slides from Tzu Phang)

Defining the Genomics Landscape of Cancer



MR Stratton *et al.* *Nature* **458**, 719-724 (2009)



Resource

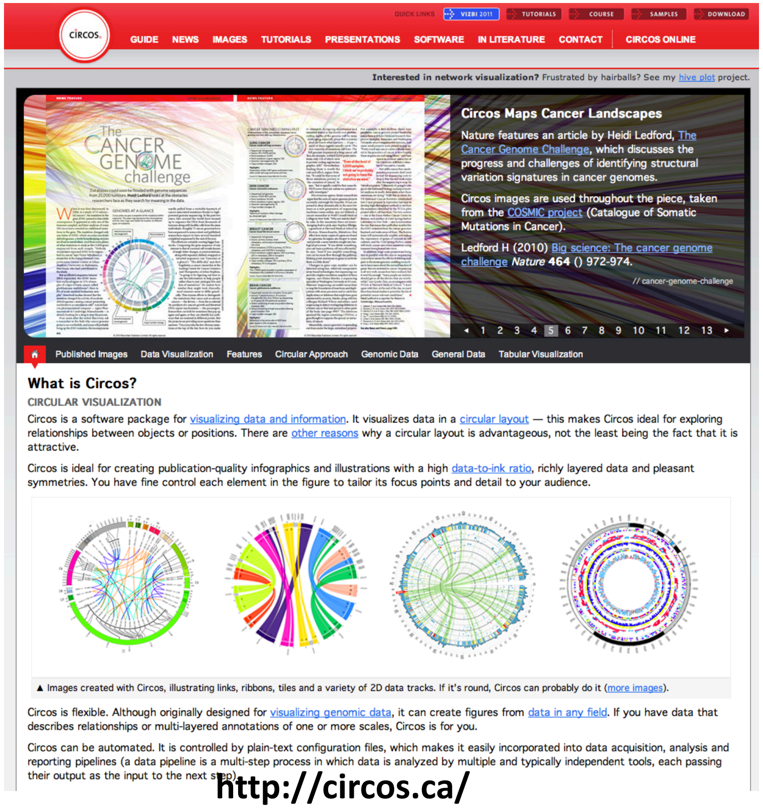
Circos: An information aesthetic for comparative genomics

Martin Krzywinski,[1,3] Jacqueline Schein,[1] İnanç Birol,[1] Joseph Connors,[2] Randy Gascoyne,[2] Doug Horsman,[2] Steven J. Jones,[1] and Marco A. Marra[1]

[1]*Canada's Michael Smith Genome Sciences Center, Vancouver, British Columbia V5Z 4S6, Canada;* [2]*British Columbia Cancer Research Center, British Columbia Cancer Agency, Vancouver, British Columbia V5Z 1L3, Canada*

We created a visualization tool called Circos to facilitate the identification and analysis of similarities and differences arising from comparisons of genomes. Our tool is effective in displaying variation in genome structure and, generally, any other kind of positional relationships between genomic intervals. Such data are routinely produced by sequence alignments, hybridization arrays, genome mapping, and genotyping studies. Circos uses a circular ideogram layout to facilitate the display of relationships between pairs of positions by the use of ribbons, which encode the position, size, and orientation of related genomic elements. Circos is capable of displaying data as scatter, line, and histogram plots, heat maps, tiles, connectors, and text. Bitmap or vector images can be created from GFF-style data inputs and hierarchical configuration files, which can be easily generated by automated tools, making Circos suitable for rapid deployment in data analysis and reporting pipelines.

[Supplemental material is available online at http://www.genome.org. Circos is licensed under GPL and available at http://mkweb.bcgsc.ca/circos. An interactive online version of Circos designed to visualize tabular data is available at http://mkweb.bcgsc.ca/circos/tableviewer.]



Interested in network visualization? Frustrated by hairballs? See my hive plot project.

Circos Maps Cancer Landscapes

Nature features an article by Heidi Ledford, The Cancer Genome Challenge, which discusses the progress and challenges of identifying structural variation signatures in cancer genomes.

Circos images are used throughout the piece, taken from the COSMIC project (Catalogue of Somatic Mutations in Cancer).

Ledford H (2010) Big science: The cancer genome challenge *Nature* **464** () 972-974.

// cancer-genome-challenge

What is Circos?

CIRCULAR VISUALIZATION

Circos is a software package for visualizing data and information. It visualizes data in a circular layout — this makes Circos ideal for exploring relationships between objects or positions. There are other reasons why a circular layout is advantageous, not the least being the fact that it is attractive.

Circos is ideal for creating publication-quality infographics and illustrations with a high data-to-ink ratio, richly layered data and pleasant symmetries. You have fine control each element in the figure to tailor its focus points and detail to your audience.

▲ Images created with Circos, illustrating links, ribbons, tiles and a variety of 2D data tracks. If it's round, Circos can probably do it (more images).

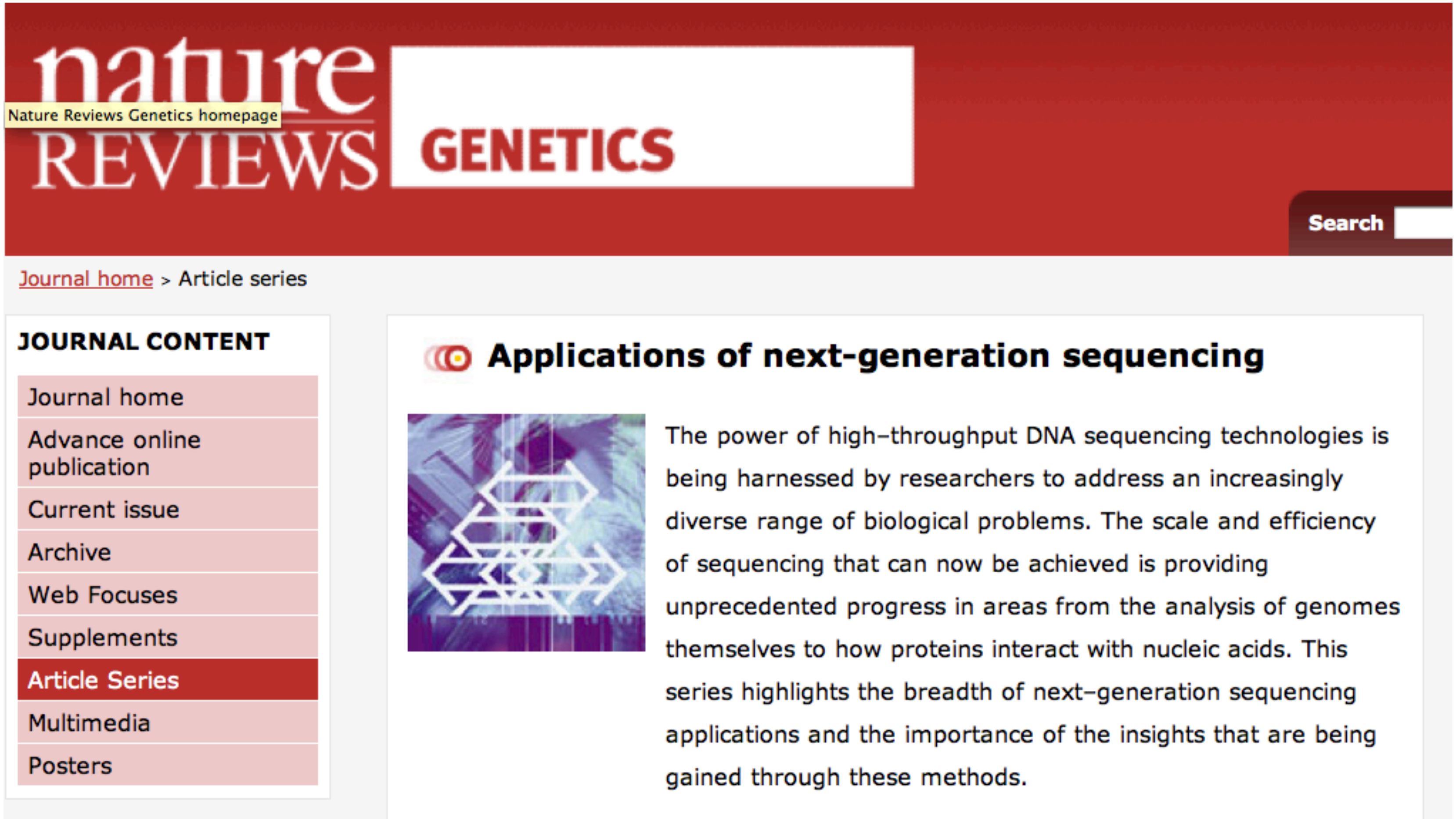
Circos is flexible. Although originally designed for visualizing genomic data, it can create figures from data in any field. If you have data that describes relationships or multi-layered annotations of one or more scales, Circos is for you.

Circos can be automated. It is controlled by plain-text configuration files, which makes it easily incorporated into data acquisition, analysis and reporting pipelines (a data pipeline is a multi-step process in which data is analyzed by multiple and typically independent tools, each passing their output as the input to the next step).

http://circos.ca/

Nature Reviews Genetics

http://www.nature.com/nrg/series/nextgeneration/index.html



Applications of next-generation sequencing

The power of high-throughput DNA sequencing technologies is being harnessed by researchers to address an increasingly diverse range of biological problems. The scale and efficiency of sequencing that can now be achieved is providing unprecedented progress in areas from the analysis of genomes themselves to how proteins interact with nucleic acids. This series highlights the breadth of next-generation sequencing applications and the importance of the insights that are being gained through these methods.

http://www.nature.com/nrg/series/nextgeneration/index.html

Take Home Message

- NGS is a powerful technology to generate single base resolution for quantifying gene expression, detecting SNP (and other mutations), specifying TF/RNA/Protein-DNA interactions and methylation (globally)
- Innovative bioinformatics tools have been developed to analyze and interpret these massive “omics” data